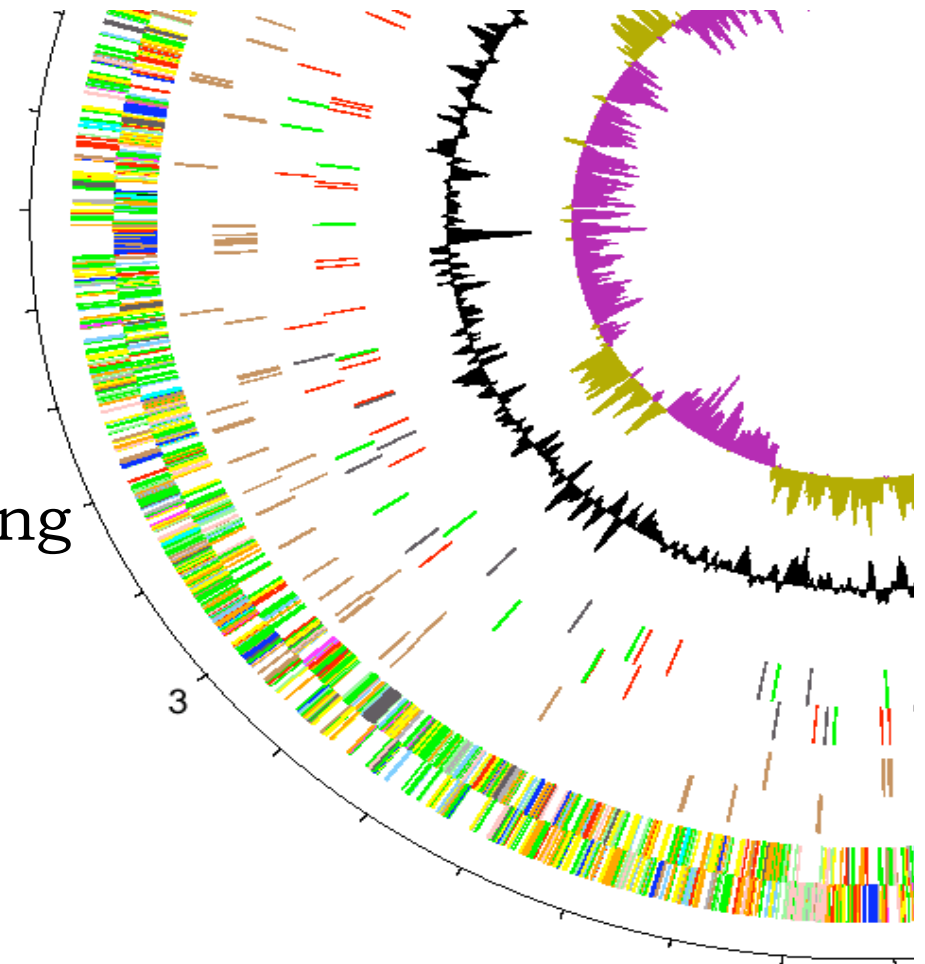
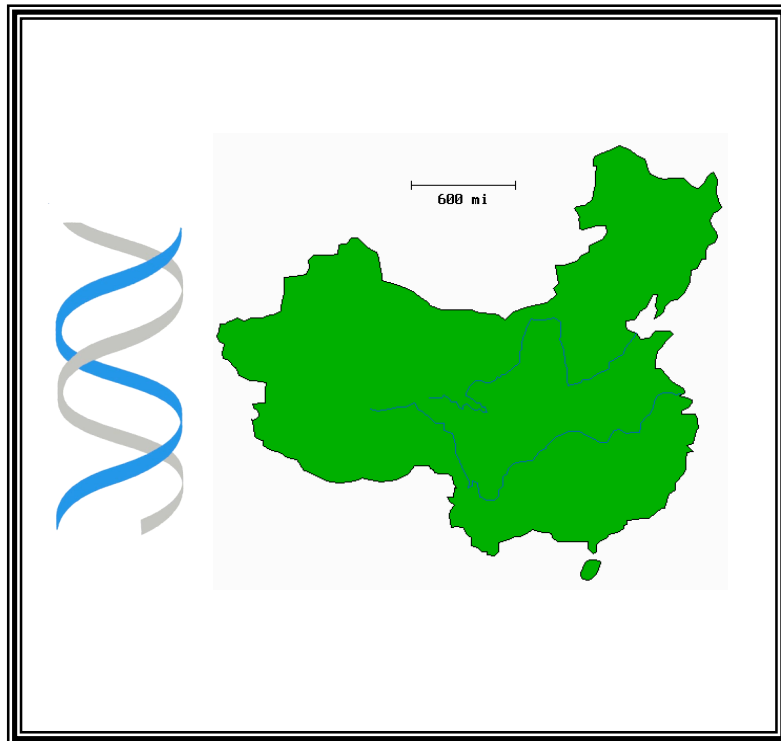


Session 2: Sequence analysis and Salmonella Genome sequencing



Nicholas R. Thomson
Wellcome Trust Sanger Institute,
Cambridge
UK

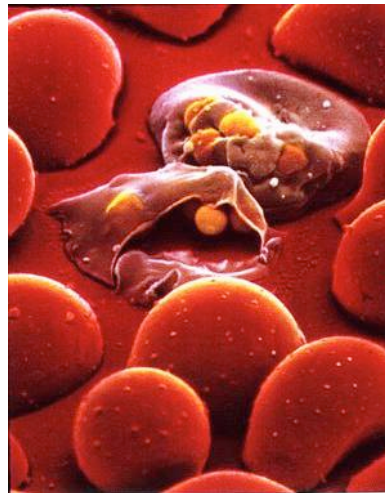


Pathogen Sequencing Unit

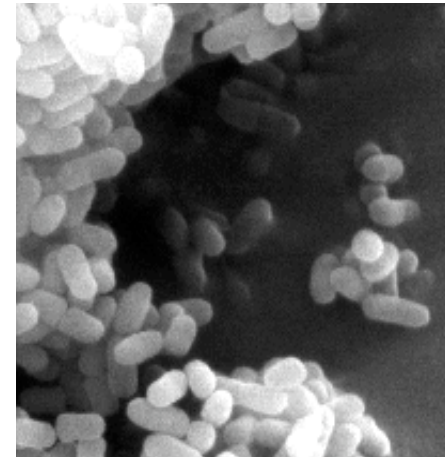
The Pathogen Group is funded by the Beowulf Genomics Initiative to sequence the genomes of a wide range of small eukaryotes and microbes



Yeasts and Fungi:
Saccharomyces
Aspergillus



Protozoa:
Plasmodium falciparum
Leishmania



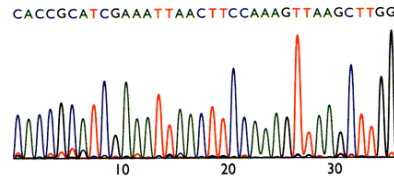
Bacteria:
M. tuberculosis
M. leprae
Y. pestis
S. typhi



Sequencing Projects

Shotgun

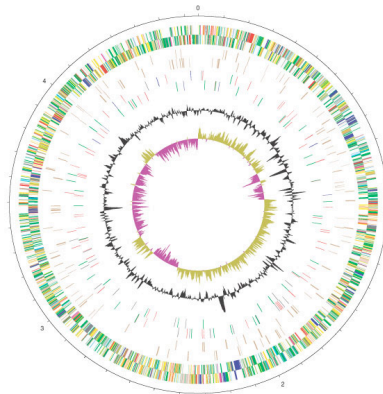
Genome sequencing Finishing



Complete Sequence

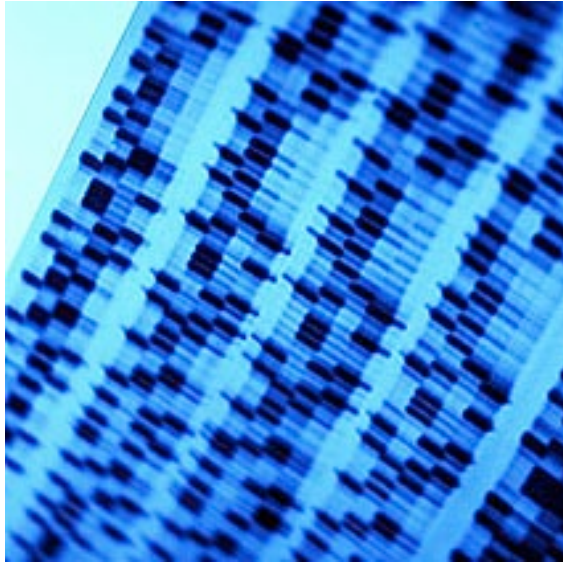
tgttgatctctgtttttttttttttttttttttttccatcaaaaattagtgaaaacata
cagctgacatcttaaatgactgttccacaggtttgtggagacaaaacagctgttat
attttttttccacaggtttgtggaaacagaagattttatgtatgagattctagaat
attccacagaggaacatgtagtatacagaatgaacaaattttttggacagactgtt
gaatctgcagctcagatcaaaaacagacattgaattttttgtctgacgccct
ctatcaagctgcagacagatttgcacattctatagctcaatgaagaacatttt
tgggaaaacacatttcaagatgtattctctacattttgttggaatttataacgctca
attctctgtacatttttggaaagagacaaatgactgagcaacaaacagacacaa
aatcaaaaactcaagcagcaacgtcaattcttctgactctcattcaatttaaaa
tcgaattattgtttgaaacatttcaaaagagatgaanaactgttgggtgtctgt
ctaatcaagctgacataactctcagacacacataactctctgtttttgggtgac
cctggctttggaaaacaaatcttaaaatgtattgtaattctgcatactagaana
cnaatgctcgaattaaatatacagctggaacatttataatgattttctttatcatt
attctctgcagacattcagctggaagaaaanaattctgaatttagatttttga
tctctgacattctacatactgacacacacacacacacacacacacacacacac
ttatctttttatgacactcaattatacaaaaacattctgcacaaaagcagctca
cagaacatctcaatgattagaagactgattgattcagcttttaaaagggttaaaa
gtcaattctacacacctctgatttggaaacacagctgctatttgacaatcaaatct
gaataatcatttttttctccagacacacatggaattttgtctgctcatttgatct
aattgcagagattttagaagctgcttaaaagattatgattgcttattttcaaaa
attgacacatttctgtatgactctgcaggaagattctgcacggcaagaagattg
ctcaanaatgacatttccactcagaagattccaagcgaagtggaaaattttcaagt
tctacgtccaagaatgaatcattcaaacgaacaaatattttttgcagcaacac
gtgactctgttttttagacagttgaatgacagatcagacattctcctaatttggaaagaa
tttgttgcagagacacattcaaacgattctcagctctaatcaatcaaaaattgac

Genome annotation and analysis



Published genome & Database entry

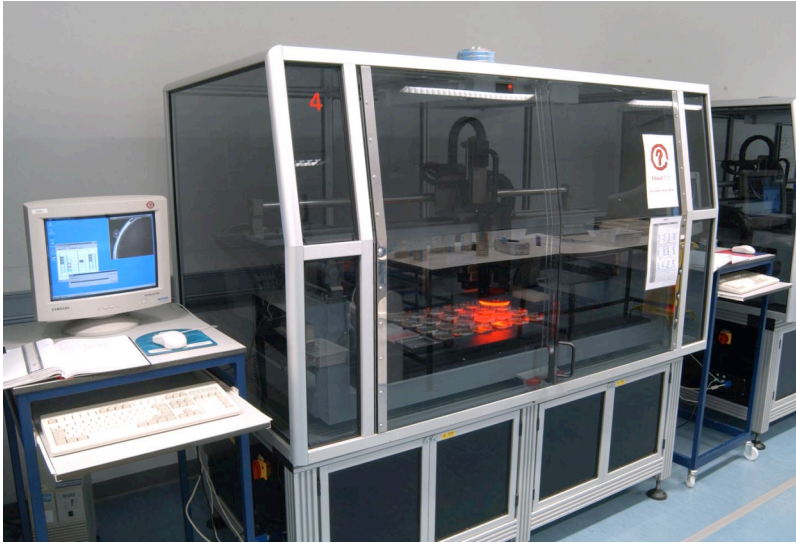




1. High throughput genome sequencing
2. Developments in technology
3. Sequence analysis
4. Comparative Genomics:
 - Core and accessory genome
 - Pathogenomics



Levels of automation

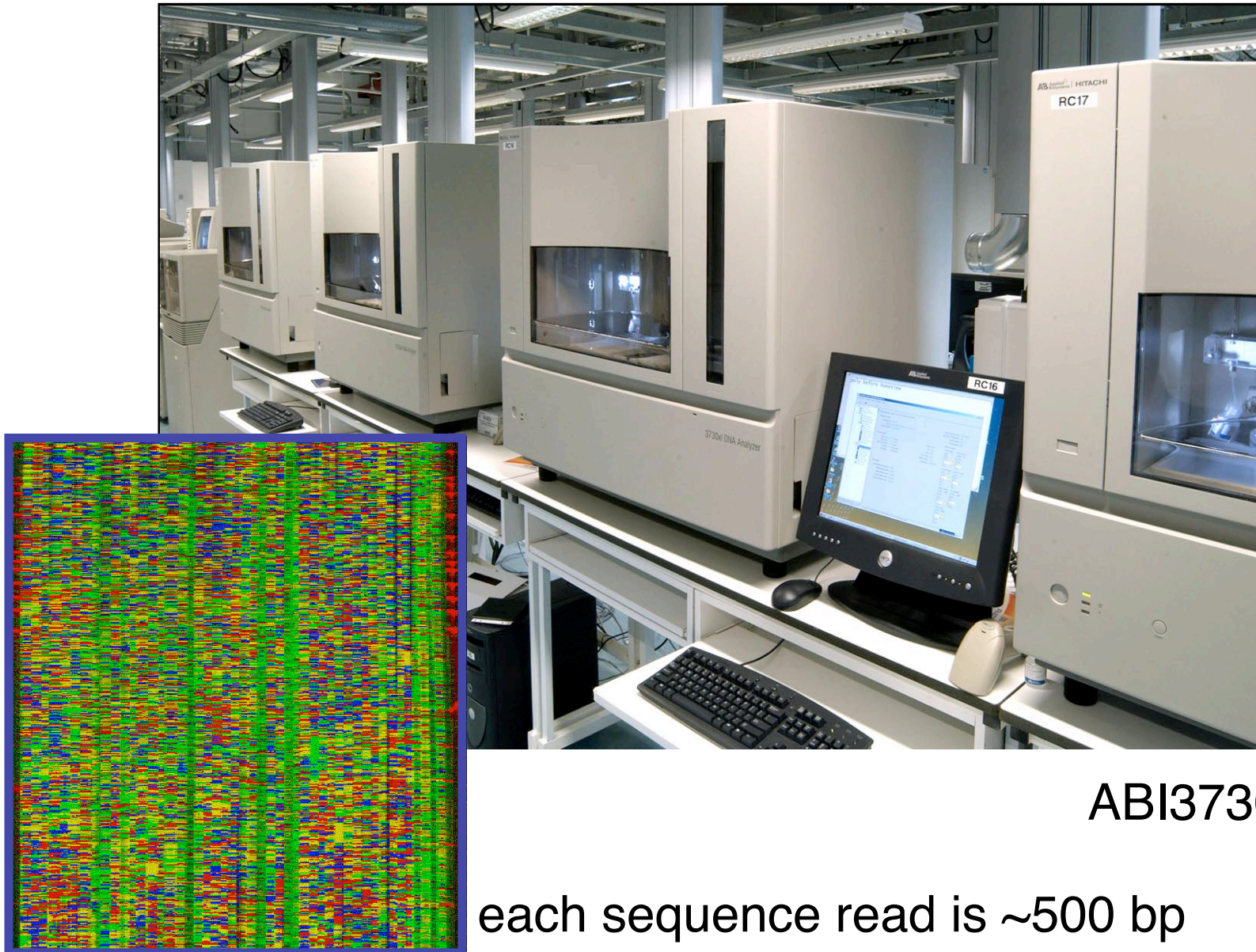


Colony picking robots





Raw sequence data



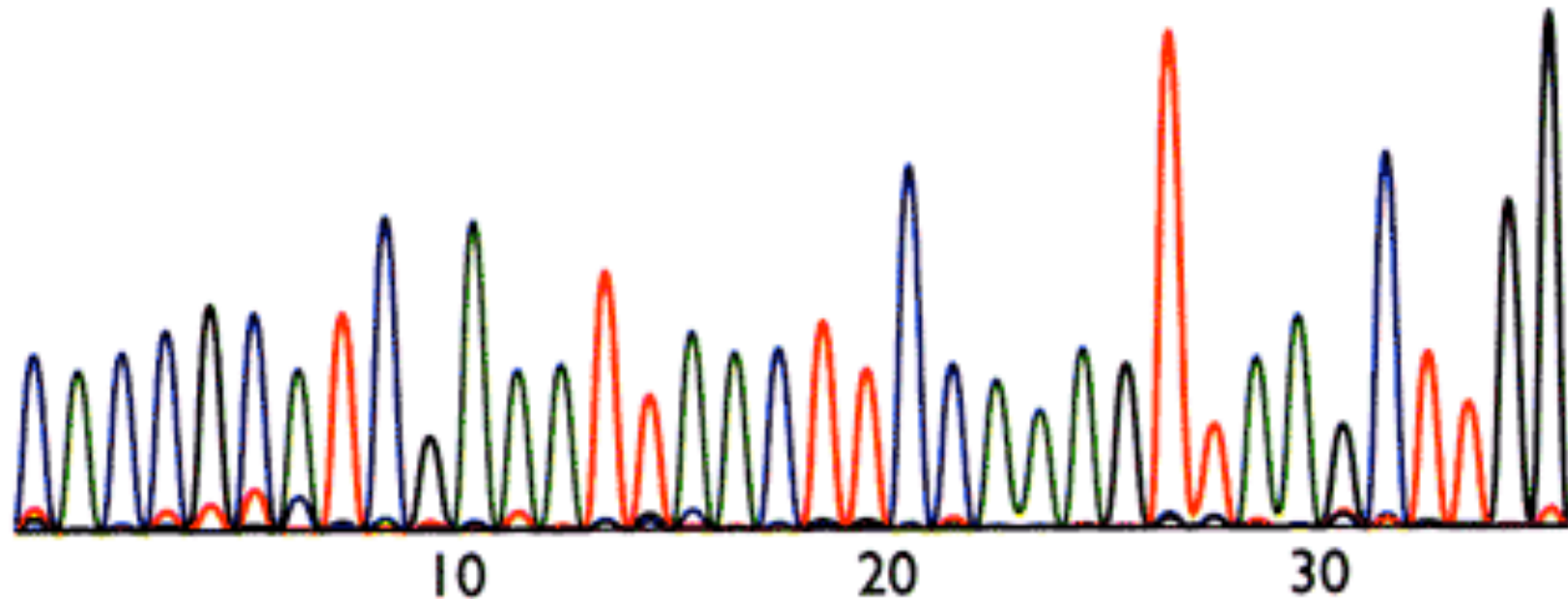
ABI3730

each sequence read is ~500 bp



Automated sequencing

CACCGCATCGAAATTAACTTCCAAAGTTAAGCTTGG

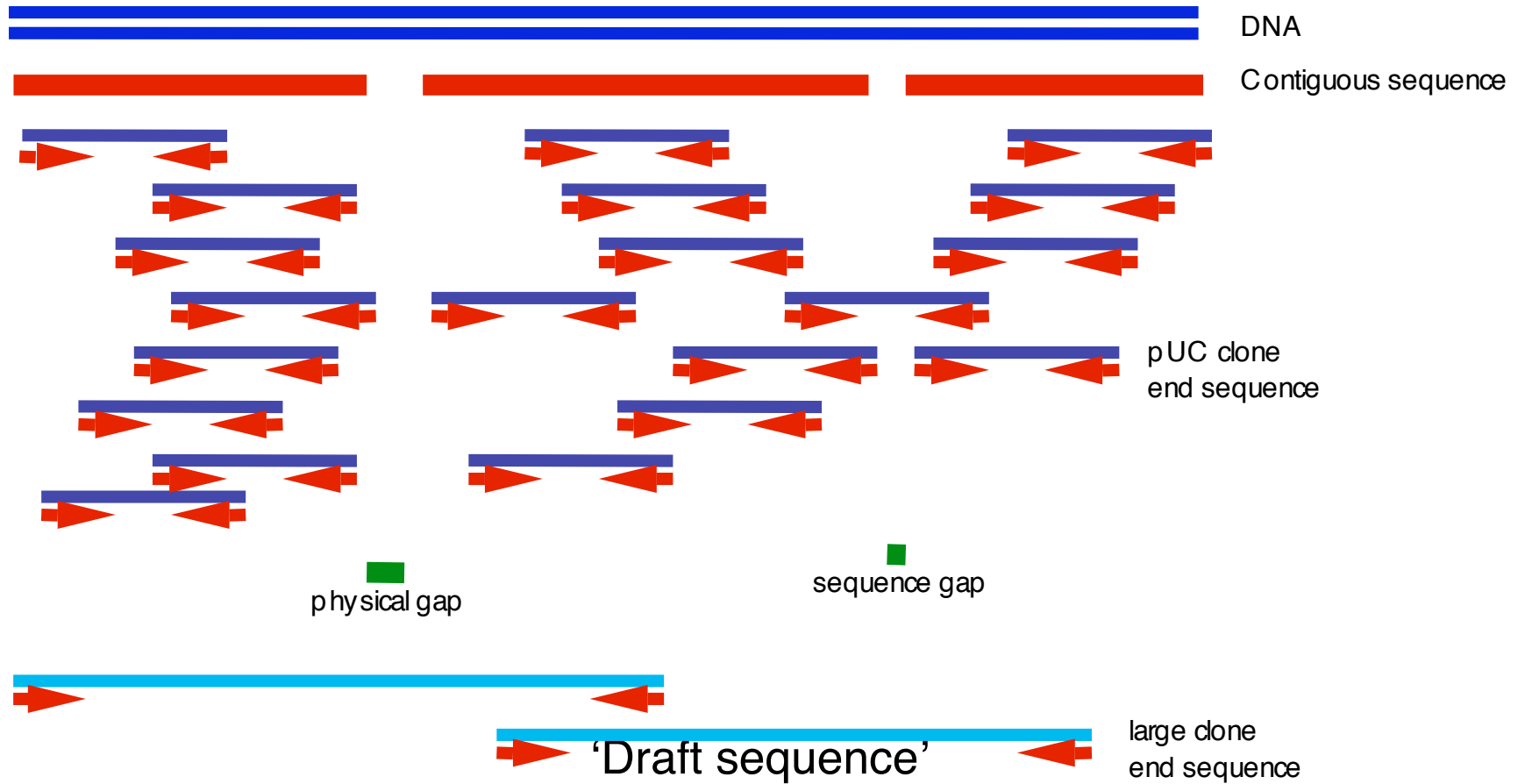


Each ABI 3700 reads 96 DNA sequences at once, producing a “trace” like this one for each Sequence.

Sequencing strategy and assembly



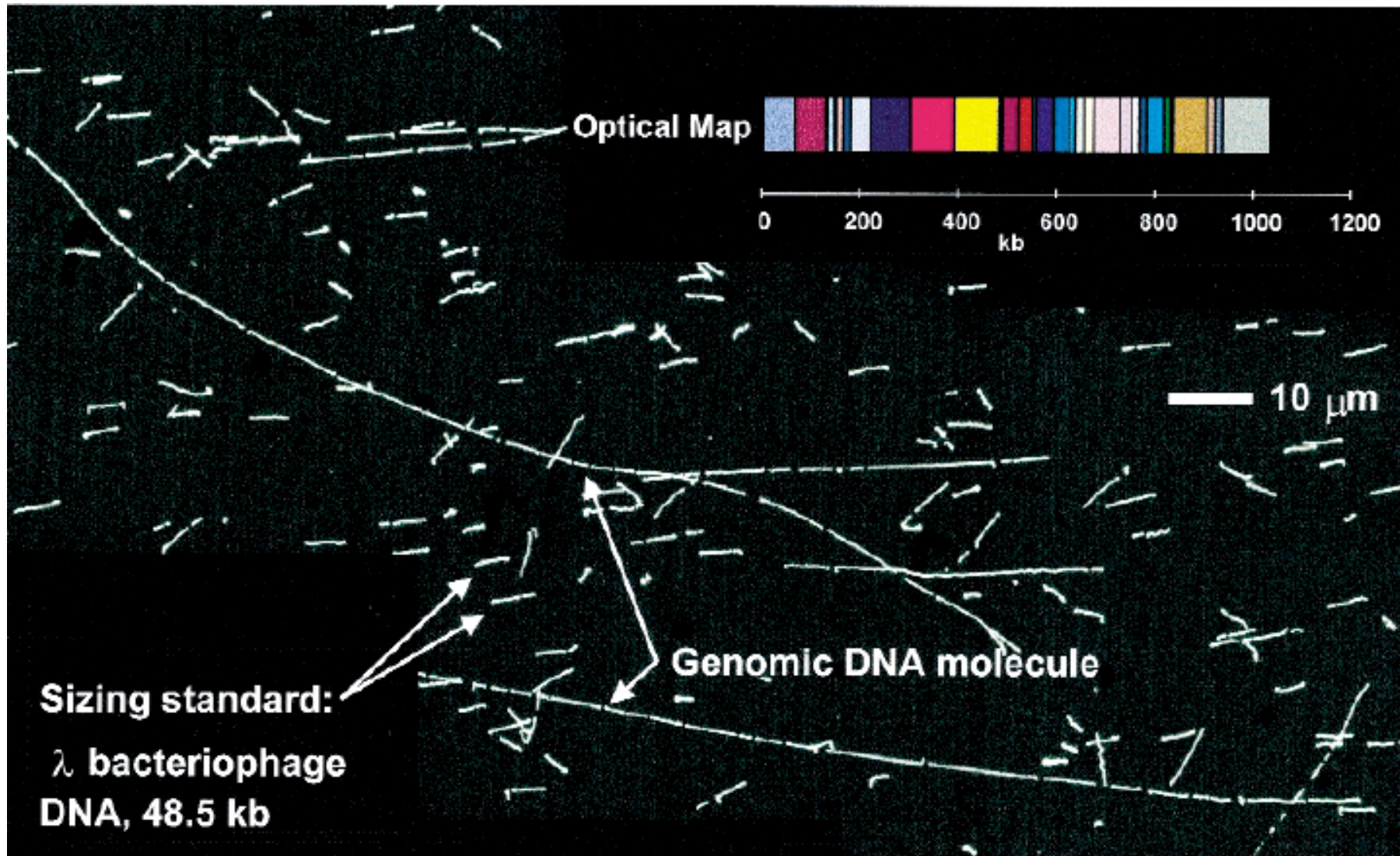
Shotgun sequencing – strategy



Order of contigs?

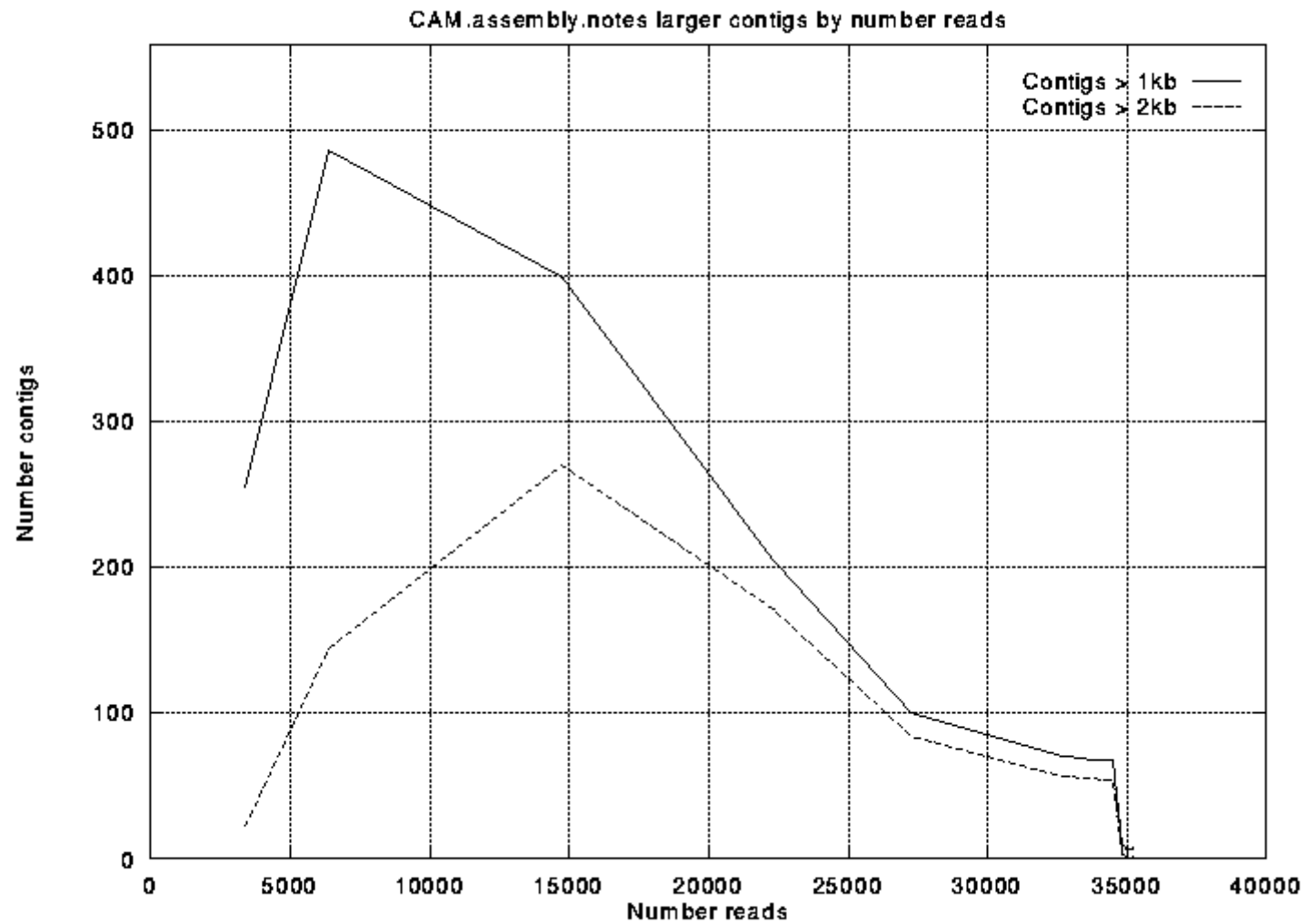
95% coverage, 4-5x depth.
Finished sequence: 100% coverage, 10x depth.

Optical Mapping



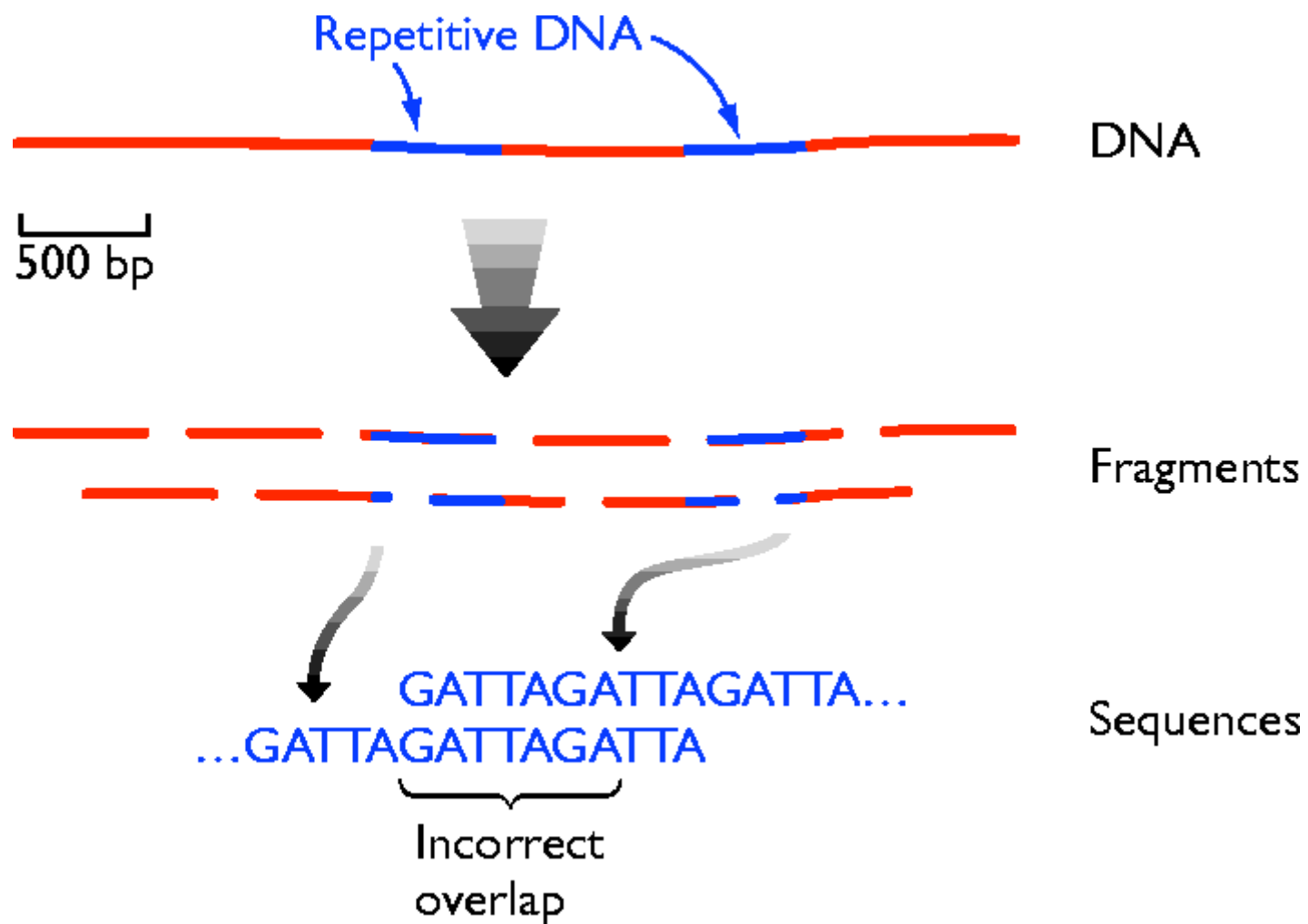


Shotgun assembly - *Campylobacter jejuni*



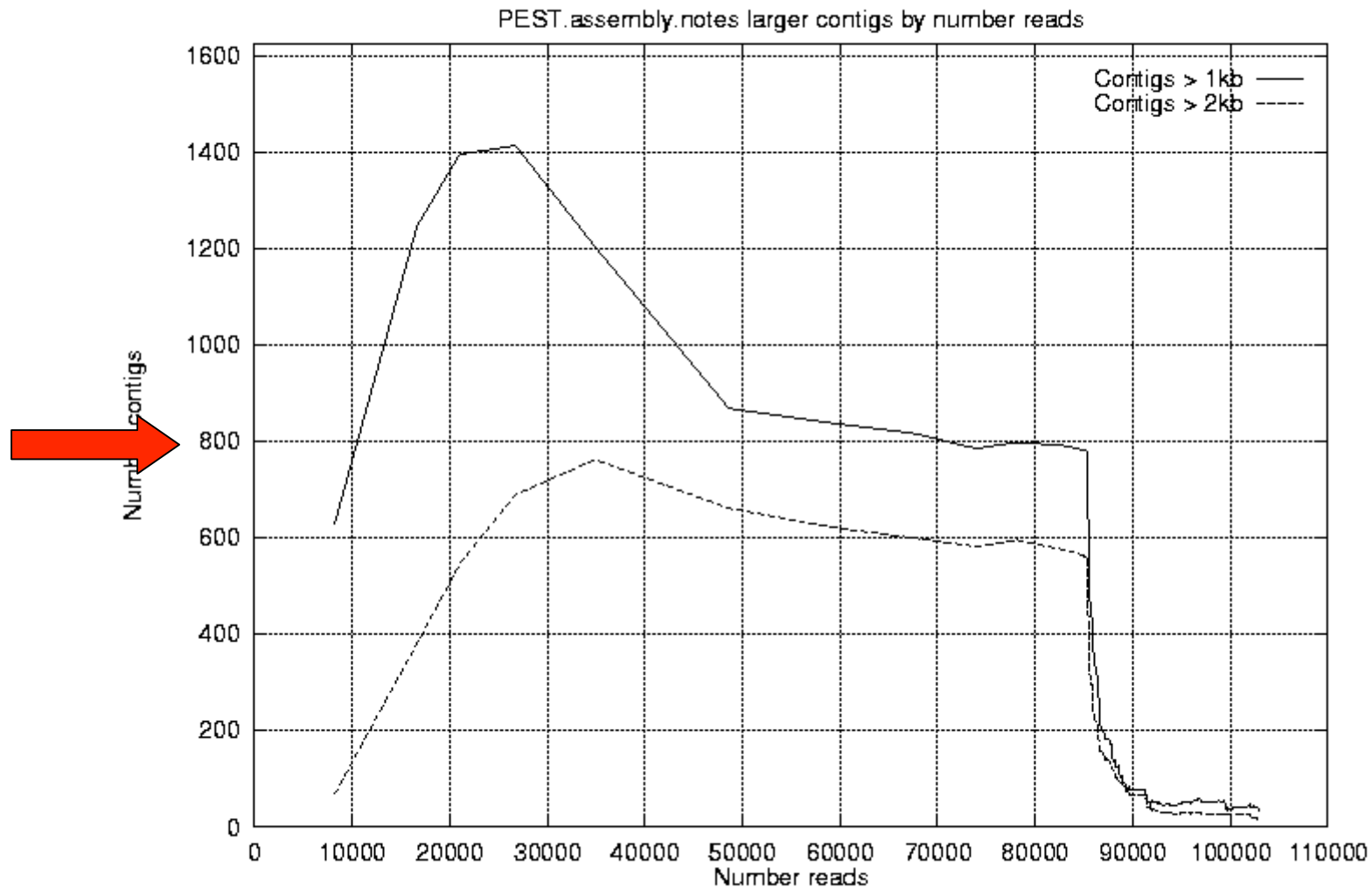


Repeats!!!





Shotgun assembly - *Yersinia pestis*

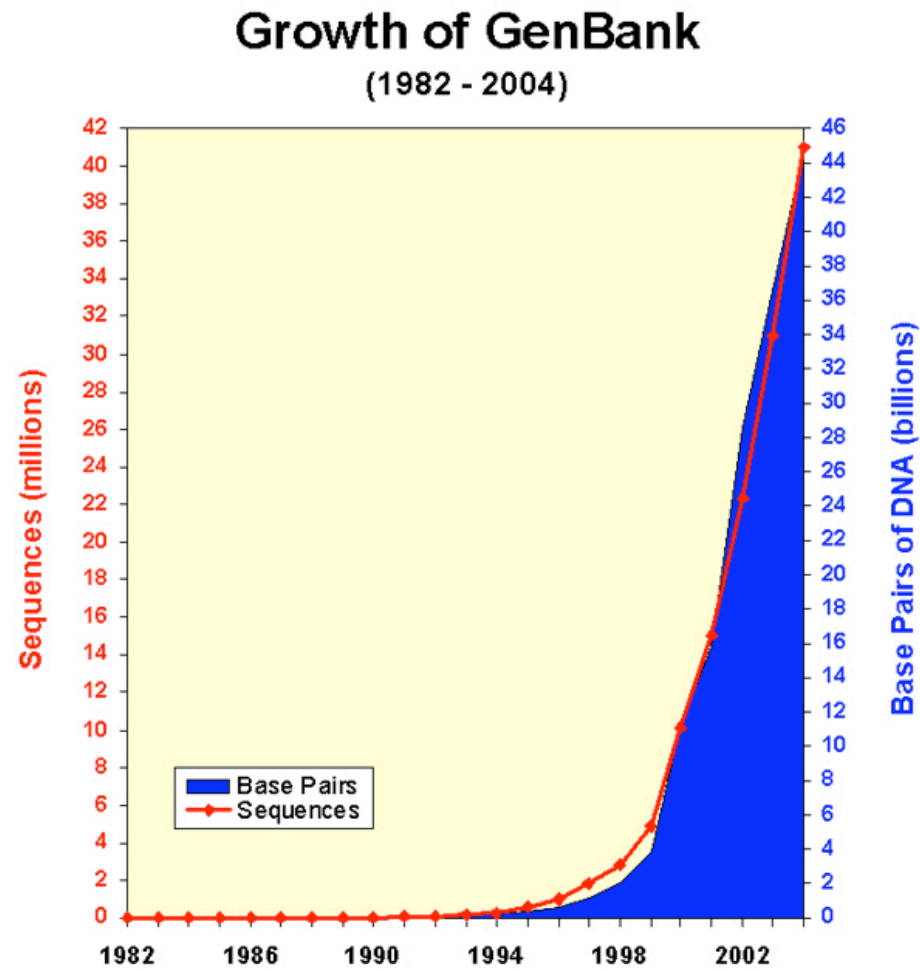


PSU Bacterial Genomes.

>250 in the databases

<i>Bacteriovorax marinus</i>	Finished	<i>Photorhabdus asymbiotica</i>	Gap closure
<i>Bacteroides fragilis</i> NCTC9343	Published	<i>Proteus mirabilis</i>	Finished
<i>Bacteroides fragilis</i> 638R	Finished	<i>Pseudomonas fluorescens</i>	Finished
<i>Bordetella avium</i>	Finished	<i>Rhizobium leguminosarum</i>	Finished
<i>Bordetella bronchiseptica</i>	Published	<i>Rhodococcus equi</i>	Shotgun in progress
<i>Bordetella parapertussis</i>	Published	<i>Salmonella bongori</i>	Finished
<i>Bordetella pertussis</i>	Published	<i>Salmonella enteritidis</i>	Finished
<i>Burkholderia cenocepacia</i>	Finished	<i>Salmonella gallinarum</i>	Gap closure
<i>Burkholderia pseudomallei</i>	Published	<i>Salmonella paratyphi</i> A	Gap closure
<i>Campylobacter jejuni</i>	Published	<i>Salmonella typhimurium</i> DT104	Finished
<i>Chlamydia trachomatis</i> Jali	Finishing	<i>Salmonella typhimurium</i> SL1344	Gap closure
<i>Chlamydia trachomatis</i> L2	Finished	<i>Salmonella typhi</i>	Published
<i>Chlamydophila abortus</i>	Published	<i>Serratia marcescens</i>	Finished
<i>Citrobacter rodentium</i>	Finishing	<i>Shigella dysenteriae</i>	Gap closure
<i>Clavibacter michiganensis</i>	Finished	<i>Shigella sonnei</i>	Finished
<i>Clostridium botulinum</i>	Finished	<i>Staphylococcus aureus</i> MRSA252	Published
<i>Clostridium difficile</i>	Finished	<i>Staphylococcus aureus</i> MSSA476	Published
<i>Corynebacterium diphtheriae</i>	Published	<i>Staphylococcus aureus</i> (X2)	Funded
<i>Ehrlichia ruminantium</i>	Published	<i>Stenotrophomonas maltophilia</i>	Finished
<i>Erwinia amylovora</i>	Gap closure	<i>Streptococcus equi</i>	Gap closure
<i>Erwinia carotovora</i>	Published	<i>Streptococcus pneumoniae</i> (x4)	Sequencing in progress
<i>Escherichia coli</i> (x3)	Sequencing in progress	<i>Streptococcus pyogenes</i>	Finished
<i>Haemophilus influenzae</i>	Shotgun in progress	<i>Streptococcus suis</i>	Finished
<i>Haemophilus parainfluenzae</i>	Shotgun in progress	<i>Streptococcus uberis</i>	Finished
<i>Helicobacter mustelae</i>	Gap closure	<i>Streptococcus zooepidemicus</i>	Gap closure
<i>Mycobacterium bovis</i>	Published	<i>Streptomyces coelicolor</i>	Published
<i>Mycobacterium leprae</i>	Published	<i>Streptomyces scabies</i>	Finished
<i>Mycobacterium marinum</i>	Finished	<i>Tropheryma whipplei</i>	Published
<i>Mycobacterium microti</i>	Shotgun in progress	<i>Vibrio salmonicida</i>	Library testing
<i>Mycobacterium tuberculosis</i>	Published	<i>Wolbachia pipientis</i>	Gap closure
<i>Neisseria lactamica</i>	Gap closure	<i>Wolbachia endosymbiont</i>	Sequencing in progress
<i>Neisseria meningitidis</i> A	Published	<i>Yersinia enterocolitica</i>	Finished
<i>Neisseria meningitidis</i> C	Finished	<i>Yersinia pestis</i>	Published

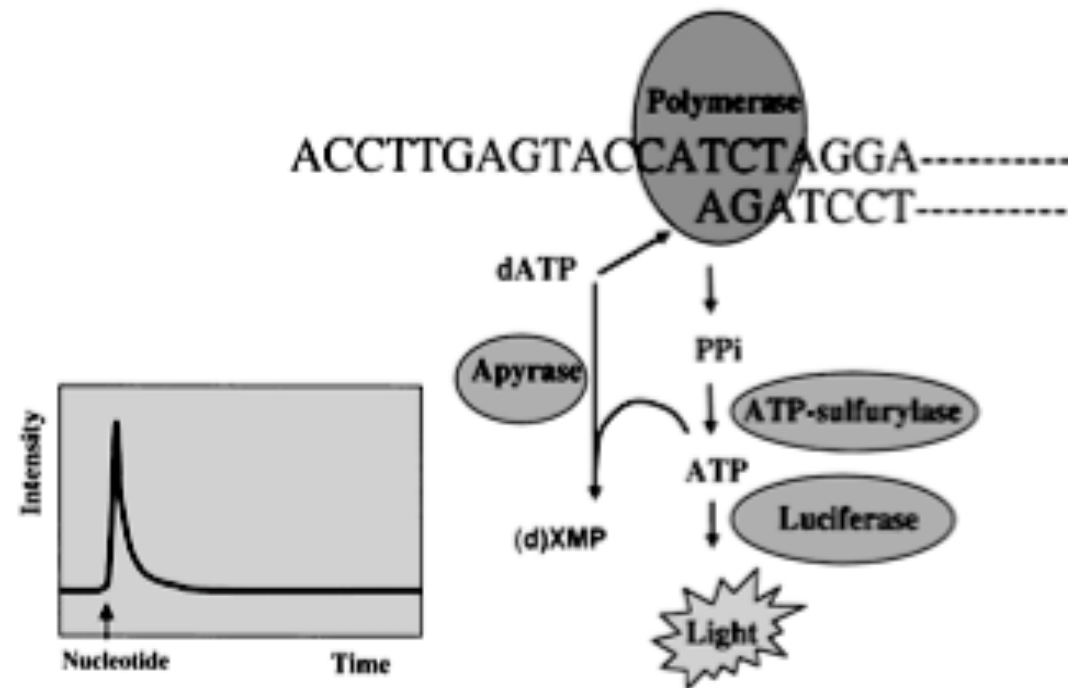
The fruits of sequencing



What will the next 10 years hold?

Developments in Technology

Pyrosequencing



Developments in Technology



- New Technologies
 - 454 sequencing
 - Clonal amplification on beads
 - Pico titre plate (1.6 M wells)
 - Sequencing-by-synthesis
 - Chemiluminescent detection
 - No cloning required
- Increased performance
 - 20,000,000 bp per run (4.5 hours)
 - 2 Mb genome, 10x coverage
- Current performance (ABI 3730)
 - 48,000 bp per run

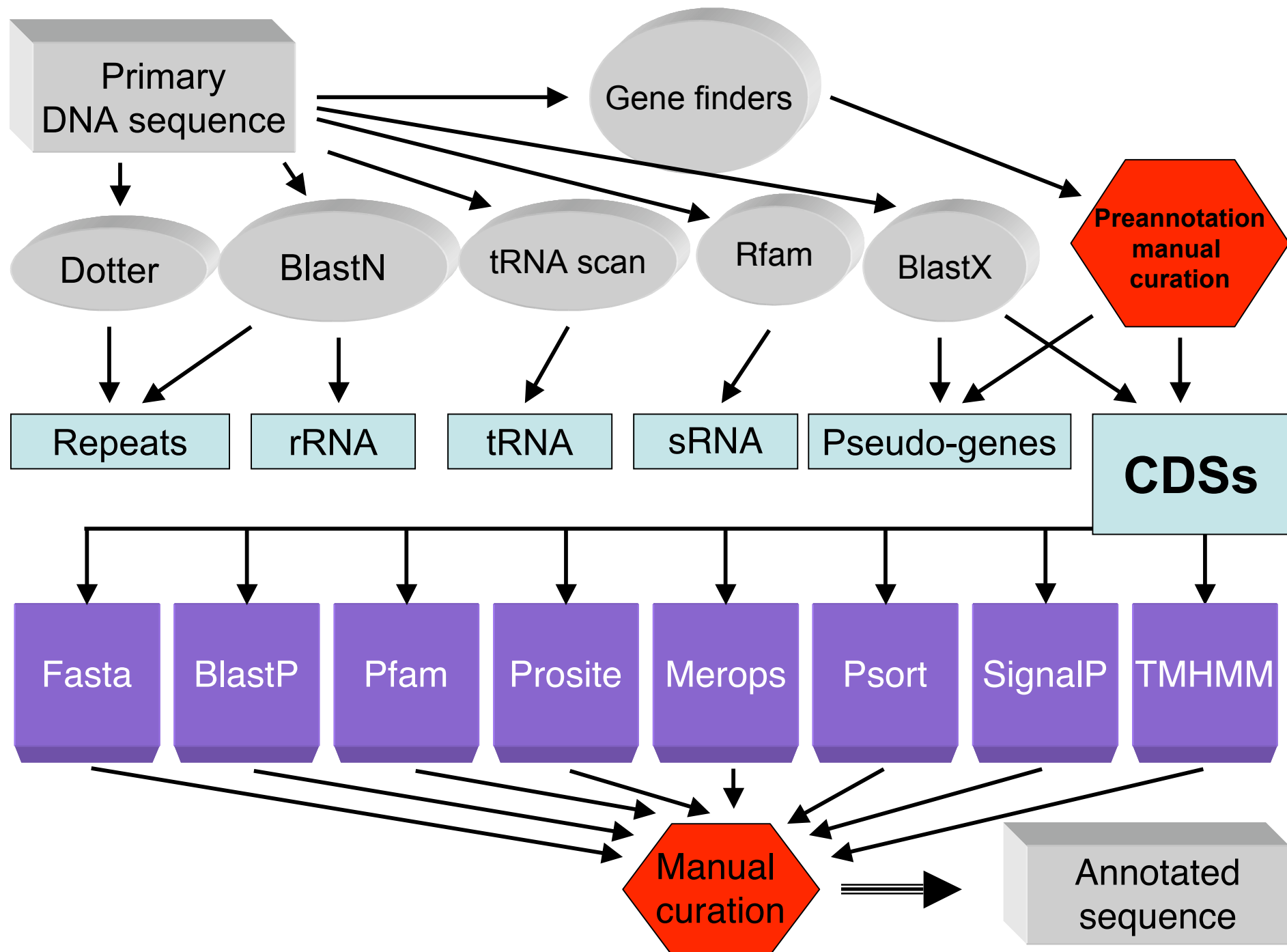
Annotation strategy

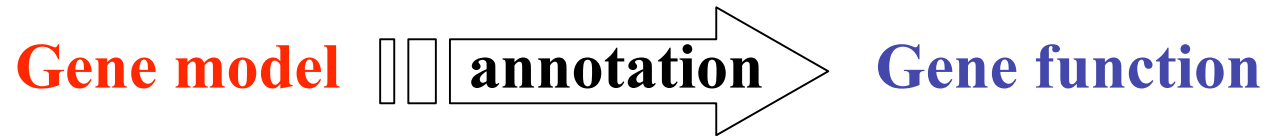
1. Overview

2. Database searches

- ***BLAST/FASTA***
- ***PROTEIN MOTIF***

3. ORTHOLOGUE/PARALOGUE





Key WWW-based resources to be covered in this module:

SRS (<http://srs.ebi.ac.uk/> or <http://srs.sanger.ac.uk/>)

Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>)

Blast searches (<http://www.ncbi.nlm.nih.gov/BLAST/>)

Fasta searches (<http://www.ebi.ac.uk/fasta33/>)

t-RNA Scan (<http://www.genetics.wustl.edu/eddy/tRNAscan-SE/>)

Rfam (<http://www.sanger.ac.uk/Software/Rfam/>)

Pfam (<http://www.sanger.ac.uk/Software/Rfam/>)

BLOCKS (<http://www.blocks.fhcrs.org/>)

TIGRfam (<http://www.tigr.org/TIGRFAMs/>)

PRINTS (<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>)

SMART (<http://www.embl-heidelberg.de/>)

PROSITE (<http://ca.expasy.org/prosite/>)

ProDom (<http://prodes.toulouse.inra.fr/prodom/current/html/home.php>)

InterPro (<http://www.ebi.ac.uk/interpro/>)

TMHMM (<http://www.cbs.dtu.dk/services/TMHMM/>)

TMPRED (http://www.ch.embnet.org/software/TMPRED_form.html)

SignalP (<http://www.cbs.dtu.dk/services/TMHMM/>)

PSORT (<http://psort.nibb.ac.jp/>)

Expasy Molecular Biology Server: (<http://ca.expasy.org/>)

Databases:

- Nucleotide / protein sequence databases:
EMBL/GenBank/DDBJ, UNIPROT, organism-specific etc.
- Protein domain / motifs:
Pfam, TIGRfam, ProSite, SMART, PRINTS, InterPro etc.
- Other specialised databases:
Rfam etc.

Sequence similarity searching:

BLAST (Basic Local Alignment Search Tool) analysis:

Nucleotide sequences:

blastn: nucleotide sequence compared to nucleotide database

blastx: nucleotide sequence translated and all 6 frame translations compared to protein database

tblastn: nucleotide sequence translated and compared to translated nucleotide database (all 6 frames in both cases).

Protein sequences

blastp: protein sequence compared to protein database

tblastx: protein sequence compared to translated nucleotide database (all 6 frames).

FastA:

Provides sequence similarity and homology searching against nucleotide and protein databases using the Fasta programs. Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences.



Global

```
O05323      PCL1233 DNA FRAGMENT.                      (240 aa)
initn:  267  init1: 177  opt:  289  z-score: 358.9  E(): 1.6e-12
Smith-Waterman score: 289;    31.5% identity in 232 aa overlap

              10      20      30      40      50
STY361      MKQPEEELQETLTLEDDRAVVDYLRHHPEFFIRNAHAVEVMRVPHPVRGTVS
              ::. . . : X: . . . : . . . : . . . : . . . : . . . : . . . :
O05323 MTDQPQVPAPQPDE--SPSLPETA--AVAAYLEAHPDFFVEHEELLATMRIPHRRGDTVS
              10      20      30      40      50

              60      70      80      90     100     110
STY361 LVEWHMARARNHINVLEENMTLLMEQAHANESLFYRLHLQSRIVAADSLEMLMRFHRW
              ::. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . X .
O05323 LVEHQMKILRERNIEMRHLSHLMDVARDNDRLFDKTRRLILALMDASTLEDLVMSVEDS
              60      70      80      90     100     110

              120     130     140     150     160     170
STY361 AR-DLGLAGATLRLFPDRWRLGAP-SRYTHLALNRQAFEPLRIQLGQSQHLYGLPLNGPE
              : . . . . . : . . . . . : . . . . . : . . . . . : . . . . . : . . . . . :
O05323 LRQDFQVPFVSLILFGDN---AMPVGRWVTHAEAQTAIGGL---LTEDKSVSGSLREHE
              120     130     140     150     160

              180     190     200     210     220
STY361 LLVVLPEA--KAVGSMGSDGDLGVILFSSRDPHHYQPGQGTQLLQEIALLMPELL
              : . . : . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
O05323 LDPLFGEEQKQIGSTAVVAIAHQGLHGVLAIASRDPQHYKSSVGTFLFLSVIAEVTGRVL
              170     180     190     200     210     220

              230
STY361 ERWIKRV
              :
O05323 PRVAGSLRSVR
              230     240
```

FASTA

Local

```
>TR:O31086 O31086 ORF235 (FRAGMENT). [0]
      Length = 63

Score = 204 (71.8 bits), Expect = 1.5e-16, P = 1.5e-16
Identities = 38/62 (61%), Positives = 50/62 (80%)

Query:   176 PEAKAVG---SVAMSMGSDGDLGVILFSSRDPHHYQPGQGTQLLQEIALLMPELLERWI 232
          P+AK +G   SVA+SM+G DG+LG+++FSSRD  HYQ G GT +L ++A MLPELLERWI
Sbjct:    1  PQAKQIGQIGSVALSMLGDDGELGMVIFSSRDTQHYYQQGMGTVMNLNQLARMLPELLERWI 60

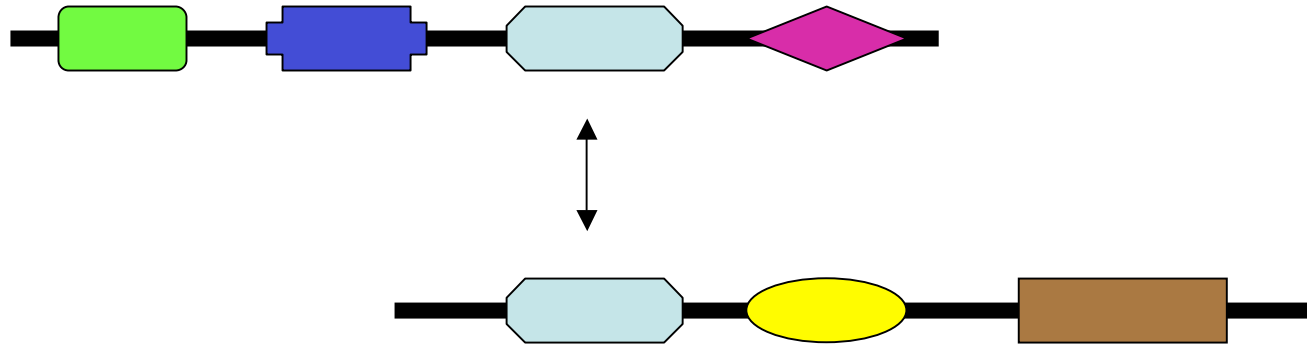
Query:   233 KR 234
          +R
Sbjct:    61 ER 62
```

BLAST

Global alignments can be more informative and trustworthy when looking at modular proteins or multifunctional proteins.

Domain problems:

Matches between similar functional domains in otherwise different proteins can lead to incorrect transfer of annotation




NCBI BLAST Home Page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Links AmiGO GO Consortium Google Parasite-GO

Address http://www.ncbi.nlm.nih.gov/BLAST/

Back Forward Stop Home Search Favorites History



NCBI

PubMed Entrez **BLAST** OMIM Taxonomy Structure

NCBI

SITE MAP

BLAST info
BLAST overview

Frequently Asked Questions

BLAST Program Selection Guide

Release 4.2.0a **NEW**

Description of BLAST Services

Subscribe to BLAST-Announce

New/Noteworthy

BLAST course

BLAST tutorial

BLAST references

URL API
documentation
HTML format

PDF format

PostScript format

BLAST

What's NEW in BLAST®

NEW March 5th 2002: New database linkouts from BLAST results. Results of a BLAST search will now link sequences from the BLAST results page to the NCBI LocusLink and UniGene databases. Links to additional databases coming soon

Nucleotide BLAST ?

- Standard nucleotide-nucleotide BLAST [blastn]
- MEGABLAST
- Search for short nearly exact matches

Protein BLAST ?

- Standard protein-protein BLAST [blastp]
- PSI- and PHI-BLAST
- Search for short nearly exact matches

Translated BLAST Searches ?

- Nucleotide query - Protein db [blastx]
- Protein query - Translated db [tblastn]
- Nucleotide query - Translated db [tblastx]

Search for conserved domains ?


Done Internet

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Links AmiGO GO Consortium Google Parasite-GO

Address VERVIEW=on&END_OF_HTTPGET=Yes&SHOW_LINKOUT=yes&GET_SEQUENCE=yes Back Search Favorites History

 **NCBI** *protein-protein* **BLAST**

Nucleotide Protein Translations Retrieve results for an RID

[Search](#)

RRPRHATQQGTHPRSPMAISLPGNNMRLTNLIEESRTGNLSTDSYLSPTWMRTSNDENA
VLSLPTPASSLNIA SGVETNP TSQQITHQHRSSVVGKPN SALIMSDLTPFPASGHHYQQ
TYDDASLHFNSLADIQSTSSAQRPRIAPDL DALFDELASLDGTD R

[Set subsequence](#) From: To:

[Choose database](#) inr

[Do CD-Search](#) ☒

Now: **BLAST!** or **Reset query** **Reset all**

Options for advanced blasting

[Limit by entrez query](#) or select from: (none)

[Composition-based statistics](#) ☒

[Choose filter](#) ☒ Low complexity ☐ Mask for lookup table only ☐ Mask lower case

[Expect](#) 10

[Word Size](#) 3

Internet

Protein profiling

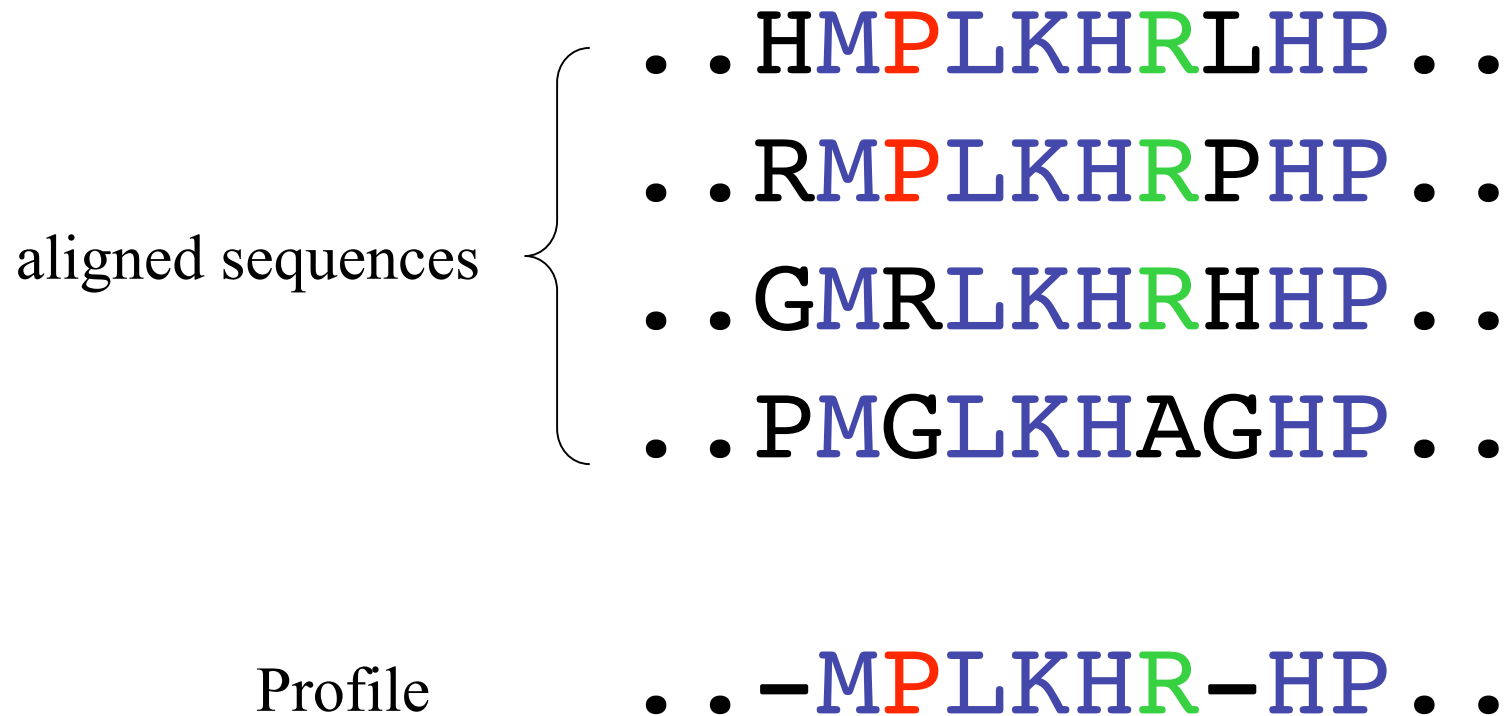
aligned sequences

• • H M P L K H P L H P • •
• • R M L L K H R P H P • •
• • G M R L K H G H H P • •
• • P M G L K H A G H P • •

Profile

• • - M - L K H - - H P • •

More sophisticated protein profiles score each amino acid in the motif



Hidden Markov Models (HMMs): The HMM is a statistical model that considers all possible combinations of matches, mismatches and gaps to generate an alignment of a set of sequences.

Profile based predictors of protein domains / motifs



Motif database in form of regular expressions. Not necessarily the whole domain.

K-x(12)-[DE] = lysine, any 12, Asparagine or glutamine.

Returns 1 or 0, i.e. very rigid and can be very inaccurate for small simple motifs



Motif search tools based on Prosite but with multiple alignment profiling



Collection of HMM's usually covering the whole domain

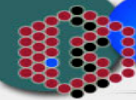


InterPro Server:

- The 'one-stop shop' for accessing all major protein databases
- InterPro provides an integrated view of the commonly used signature databases, and has an intuitive interface for text- and sequence-based searches.



InterPro: member databases


EMBL-EBI
 European Bioinformatics Institute

Get for
 Site search





[EBI Home](#)
[About EBI](#)
[Research](#)
[Services](#)
[Toolbox](#)
[Databases](#)
[Downloads](#)
[Submissions](#)

[Site Map](#)
[SRS](#)
[Start Session](#)

InterPro

- InterPro Index
- Text Search
- Sequence Search
- Databases
- Documentation
- FTP Site

InterPro Member Databases

InterPro Member Databases	
	The SWISS-PROT database consists of sequence entries. It contains high-quality annotation, is non-redundant and cross-referenced to many other databases.
	The TrEMBL database is a computer-annotated supplement to SWISS-PROT. TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database, which are not yet integrated into SWISS-PROT.
	PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.
	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains.
	PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a composite of SWISS-PROT + SP-TrEMBL. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, their full diagnostic potency deriving from the mutual context afforded by motif neighbours.
	The ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches (Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ, 1997, Nucleic Acids Res. , 25:3389-3402; Gouzy J., Corpet F. & Kahn D., 1999, Computers and Chemistry 23:333-340.) Large families are much better processed with this new procedure than with the former DOMAINER program (Sonnhammer, E.L.L. & Kahn, D., 1994, Protein Sci. , 3:482-492).
	SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signalling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa.
	TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, Hidden Markov Models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. Those entries which are "equivalogs" group homologous proteins which are conserved with respect to function.

Gene function identification

We define *homologous* proteins as those that are descended from a common ancestor, and therefore may have similar or identical functions.

We can use several tools to assist in the determination of homology including:

- statistical significance
- orthology/paralogy comparisons



Statistical significance

Statistical significance is given as “the likelihood that any given result would have occurred by chance”.

Virtually all search programs now attach a statistical value to each result, generally expressed as:

- a P-value (the probability that the match observed could have occurred by chance)

or

- an E-value (the number of results with this score expected by chance).

Statistically, figures of <0.01 are considered to be significant, although cut-offs of much lower than this are almost always used.

Statistical measures like these are strongly influenced by the size and composition of both the search sequence and the database.



Statistical significance

Statistical significance is only a tool.

Statistical significance does not necessarily equal biological significance.

Y. pestis
transcriptional
regulator *ttk*

The best scores are:

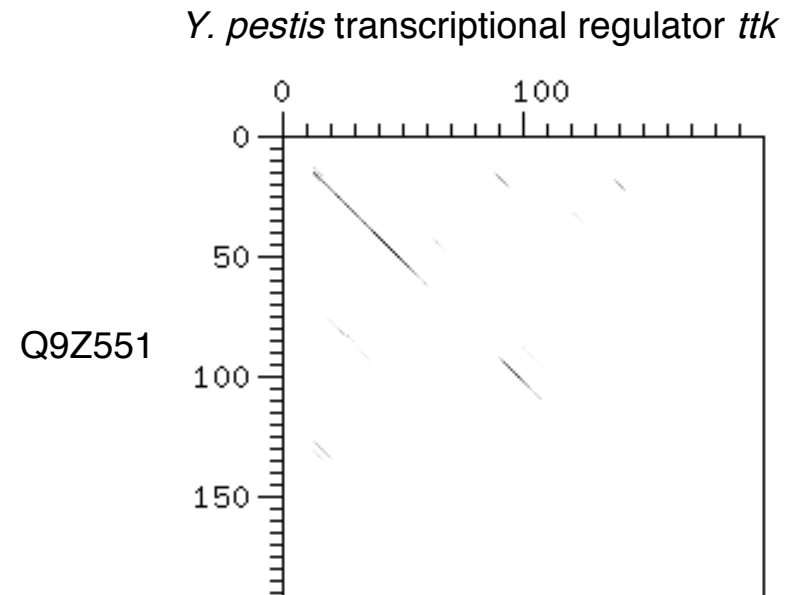
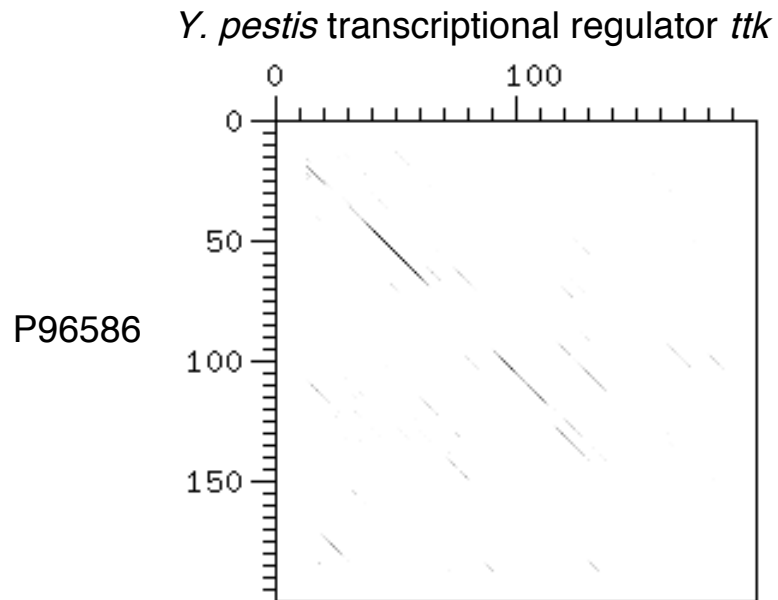
		initn	initl	opt	z-sc	E(535121)
TTK_ECOLI	TTK PROTEIN.	1085	1085	1085	1277.5	0
AAF64285	HYPOTHETICAL 24.2 KDA PROTEIN.	1061	1061	1066	1254.9	0
Q9KVD2	TRANSCRIPTIONAL REGULATOR, TETR FAMILY	856	856	868	1024.2	0
TTK_HAEIN	TTK PROTEIN HOMOLOG.	623	623	662	783.9	0
O67927	TRANSCRIPTIONAL REGULATOR (TETR/ACRR F	125	125	172	213.6	0.00023
Q9X7X6	PUTATIVE REGULATORY PROTEIN.	106	106	160	194.8	0.0026
Q9S3L4	AMTR PROTEIN.	96	96	147	183.5	0.011
Q59306	30S RIBOSOMAL PROTEIN S21.	78	78	145	182.2	0.013
Q9KIL9	F58R (FRAGMENT).	72	72	143	181.4	0.014
CAC01371	PUTATIVE TETR-FAMILY TRANSCRIPTIONAL R	72	72	143	179.1	0.019
Q9L078	PUTATIVE TETR-FAMILY REGULATORY PROTEI	128	101	142	178.1	0.022
Q9KXK1	PUTATIVE TETR-FAMILY REGULATORY PROTEI	78	51	142	176.9	0.026
CAC01492	PUTATIVE TRANSCRIPTIONAL REGULATORY PR	61	61	139	174.3	0.036
O67930	HYPOTHETICAL 22.1 KDA PROTEIN.	86	86	137	172.8	0.044
AAG04792	PROBABLE TRANSCRIPTIONAL REGULATOR.	93	93	137	172.2	0.047
Q9WZT0	TRANSCRIPTIONAL REGULATOR, TETR FAMILY	57	57	135	170.5	0.058
Q9RY76	TRANSCRIPTIONAL REGULATOR, TETR FAMILY	45	45	134	169.1	0.07
★ Q9Z551	PUTATIVE TRANSCRIPTIONAL REGULATOR.	116	116	133	168.1	0.079
★ P96856	HYPOTHETICAL 21.9 KDA PROTEIN.	75	75	133	167.9	0.082
CAC08387	PUTATIVE TETR-FAMILY TRANSCRIPTIONAL R	76	76	133	167.7	0.084



Statistical significance

Statistical significance is only a tool.

Statistical significance does not necessarily equal biological significance.



Statistical significance

Statistical significance does not necessarily equal biological significance.

Size and composition can affect statistics:

		The best scores are:	initn	initl	opt	z-sc	E(534512)
<i>Y. pestis</i> YapH protein		CAC14227 YAPH PROTEIN.	22769	22769	22769	21273.7	0
		CAC14223 YAPD PROTEIN.	9153	9153	9153	8553.6	0
		Q9JMS3 YCHA PROTEIN.	1831	773	1708	1595.5	0
		AIDA_ECOLI ADHESIN AIDA-I PRECURSOR.	1013	340	1413	1320.2	0
		Q9Z625 MISL.	1569	1020	1314	1229.6	0
		P77286 HYPOTHETICAL 50.5 KDA PROTEIN.	1358	704	1209	1136.0	0
		Q9XCJ4 SHDA.	899	321	1129	1051.8	0
		P75997 FROM BASES 1220357 TO 1232354	1017	652	1034	974.5	0
		Q52298 PLASMID PMYSH6000 VIRULENCE-ASSOCIATED	1156	428	1007	941.7	0
		CAC14226 YAPG PROTEIN.	1208	771	986	922.8	0
3705 aa, 49% Ala, Arg, Gly		CAC03614 HYPOTHETICAL 68.5 KDA PROTEIN.	950	620	931	874.1	0
		YDEK_ECOLI HYPOTHETICAL 136.5 KDA LIPOPROTEIN IN	797	392	896	836.8	0
		Q9JMS5 YCBB PROTEIN.	129	87	737	686.4	1.1e-30
	★	O50379 PPE-FAMILY PROTEIN.	210	78	725	670.4	8.3e-30
	★	O06304 HYPOTHETICAL 327.0 KDA PROTEIN PPE.	121	71	721	667.4	1.2e-29
		YFAL_ECOLI HYPOTHETICAL 131.2 KDA PROTEIN IN UBIG	574	285	631	589.5	2.7e-25
		YFAL_ECOLI HYPOTHETICAL 131.2 KDA PROTEIN IN UBIG	574	285	631	589.5	2.7e-25
		Q9XC47 OUTER MEMBRANE PROTEIN A.	190	73	607	563.7	7.3e-24
	★	O07231 RV0304C.	164	117	581	539.2	1.7e-22
	★	YZ08_MYCTU HYPOTHETICAL PE-PGRS FAMILY PROTEIN RV	164	96	563	523.3	1.3e-21
Large Ala,Arg,Gly-rich <i>M. tuberculosis</i> proteins							



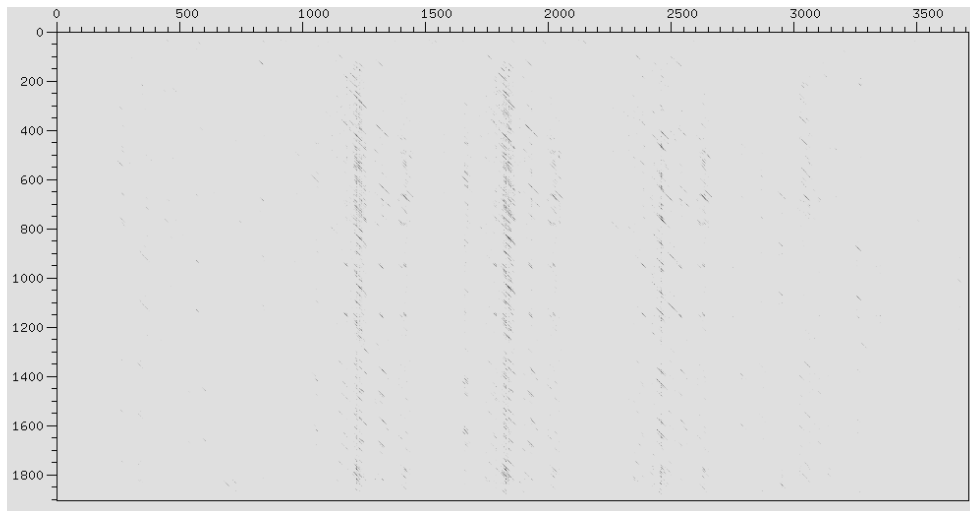
Statistical significance

Statistical significance does not necessarily equal biological significance.

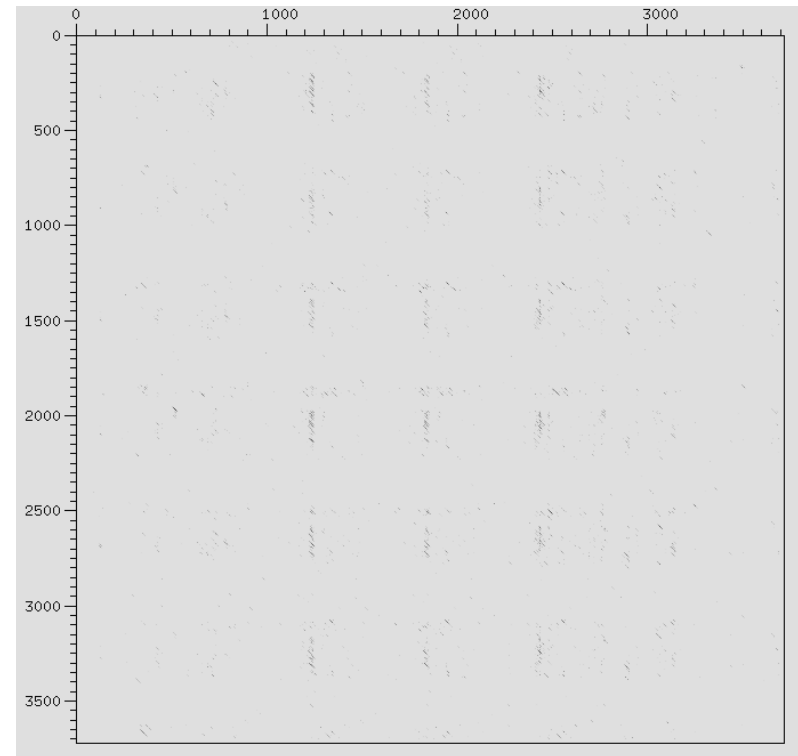
Size and composition can affect statistics:

M. tuberculosis
PPE protein

Y. pestis YapH protein



Y. pestis YapH protein



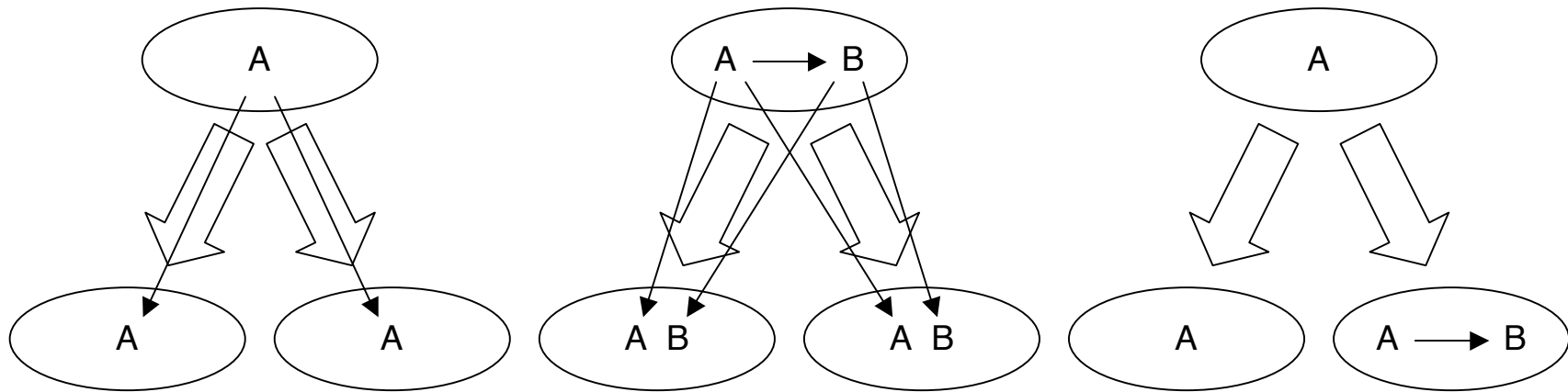
M. tuberculosis
PGRS protein



Orthology and paralogy

Orthologues can be defined as proteins in different organisms having descended from the same protein in the last common ancestor of those organisms.

Paralogues can be defined as proteins related by a duplication event at some point.



A and A are orthologues

A and A are orthologues
A and B are paralogues

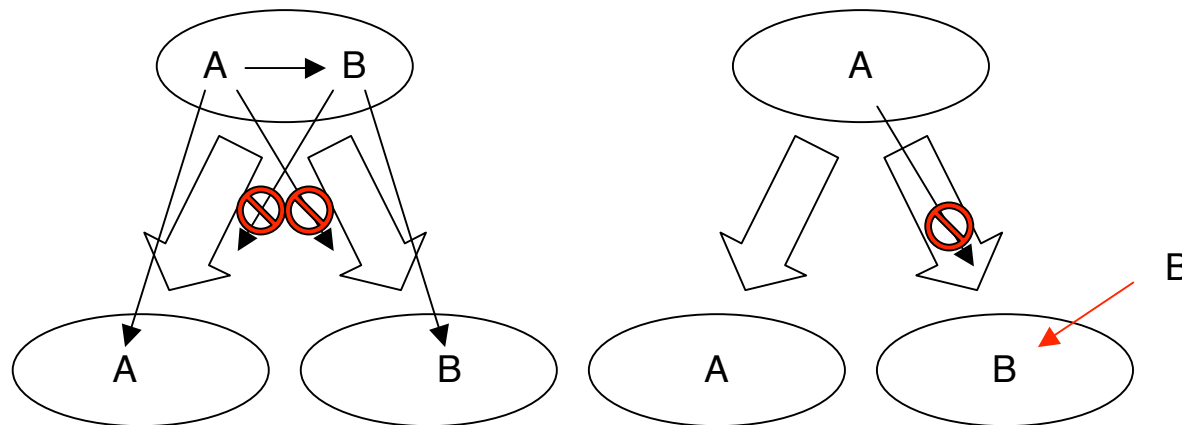


Orthology and paralogy

Orthologues are more likely to share the same function than paralogues.

Computationally we can look for orthologues by calculating reciprocal best matches between pairs of complete protein sets. This can be extended to multiple complete protein sets to form Clusters of Orthologous Groups (COGs)

This can be confounded by unequal gene loss or horizontal transfer.



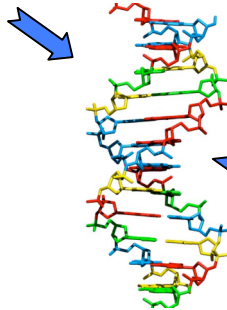
A and B may be reciprocal best matches, but are not orthologues



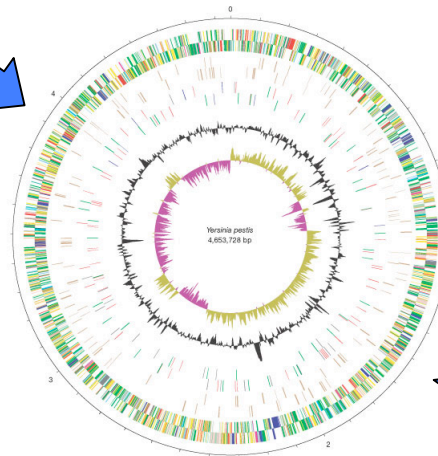
PSU Projects



Organism



Finished genome



Annotated genome

Artemis

Database entry

```
ID      AF060869      standard; DNA; PRO; 27290 BP.
XX
AC      AF060869;
XX
SV      AF060869.1
XX
DT      17-JUL-1998 (Rel. 56, Created)
DT      17-JUL-1998 (Rel. 56, Last updated, Version 1)
XX
DE      Salmonella typhimurium excision nuclease UvrA (uvrA) gene, partial cds;
DE      single-strand binding protein (ssb) gene, complete cds; tRNA-Thr gene,
DE      complete sequence; pathogenicity island SPI-4 operon, complete sequence;
DE      yjcB gene, complete cds; and yjcC gene, partial cds.
DE
DS      Salmonella typhimurium
OC      Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae;
OC      Salmonella.
XX
RN      [1]
RP      1-27290
RX      MEDLINE; 98298059.
RA      Wong K.K., McClelland M., Stillwell L.C., Sisk E.C., Thurston S.J.,
RA      Saffer J.D.
RT      "Identification and sequence analysis of a 27-kilobase chromosomal fragment
RT      containing a Salmonella pathogenicity island located at 92 minutes on the
RT      chromosome map of Salmonella enterica serovar typhimurium LT2";
RL      Infect. Immun. 66(7):3365-3371(1998).
DR      SPTREMBL: 085309; 085309.
DR      SPTREMBL: 085310; 085310.
XX
FH      Key      Location/Qualifiers
FH
FT      source      1..27290
FT                /db_xref="taxon:602"
FT                /organism="Salmonella typhimurium"
FT                /strains="LT2"
FT                /map="92 minutes"
FT      CDS        complement(1..312)
FT                /codon_start=1
FT                /db_xref="SPTREMBL:085309"
FT                /transl_table=11
FT                /genes="uvrA"
FT                /product="excision nuclease UvrA"
FT                /protein_id="AAC26637.1"
FT                /translation="MDKIEVKGARTHNKINFIIPDKLIVVTGLSGSGKSSLPDITL
FT                YREGQRRYVESLSAYARQLSLMEKPDVDHIEGLSPRAISIEQKSTSHNPRSTVGTITEI
FT                "
FT      CDS        583..1083
FT                /codon_start=1
FT                /db_xref="SPTREMBL:085310"
FT                /transl_table=11
FT                /genes="ssb"
FT                /product="single-strand binding protein"
FT                /protein_id="AAC26638.1"
FT                /translation="MILVGNPQDPEVRYMPSGGAVNHLTATSESRDKQTGEMKEQT
FT                EDHRVVMFGLAEVHGEYLRLSSQVYIEGLRTRKRTDQNCQERYTTELTSADRRVMQI
FT                LGQPKGGAPAGAHNRGLGSPQPDQPGGNGFNNGAGSRPQDSAPPSKEFPHDFDQD
FT                IPF"
FT      tRNA        1898..1970
FT                /note="putative"
FT                /product="tRNA-Thr"
XX
SQ      Sequence 27290 BP; 7965 A; 5530 C; 6661 G; 7134 T; 0 other;
gatcttcgta atagtaacca ccgtagagcg cgggttgsc gatgtcgatt tcgtgtcaat      60
tsgagtcgcg gcgcatagcc cctcaatatz gtgcgaaccc ggtttttcca tsgagcaaaa      120
aaactgcgcg gcgtaagcgg agagcgattc aacgtaacga cgtgccttc cggcatacacg      180
ttgtatttac gtatcatggc aggctgagaa gcttttcaga agagacacct tataaaataa      2520
aggcttggtt agaagacaaa atcaatagta attattgat agaattggt attcttcagg      2580
//
```

Two-character line code indicates the type of information contained in the line

Key
Qualifier

```
ID AF060869 standard; DNA; PRO; 27290 BP.
XX
AC AF060869;
XX
SV AF060869.1
XX
DT 17-JUL-1998 (Rel. 56, Created)
DT 17-JUL-1998 (Rel. 56, Last updated, Version 1)
XX
DE Salmonella typhimurium excision nuclease UvrA (uvrA) gene, partial cds;
DE single-strand binding protein (ssb) gene, complete cds; tRNA-Thr gene,
DE complete sequence; pathogenicity island SPI-4 operon, complete sequence;
DE yjcB gene, complete cds; and yjcC gene, partial cds.
DE
OS Salmonella typhimurium
OC Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae;
OC Salmonella.
XX
RN [1]
RP 1-27290
RX MEDLINE: 98298059.
RA Wong K.K., McClelland M., Stillwell L.C., Sisk E.C., Thurston S.J.,
RA Saffer J.D.;
RT "Identification and sequence analysis of a 27-kilobase chromosomal fragment
RT containing a Salmonella pathogenicity island located at 92 minutes on the
RT chromosome map of Salmonella enterica serovar typhimurium LT2";
RL Infect. Immun. 66(7):3365-3371(1998).
DR SPTREMBL: 085309; 085309.
DR SPTREMBL: 085310; 085310.
XX
FH Key Location/Qualifiers
FH
FT source 1..27290
FT /db_xref="taxon:602"
FT /organism="Salmonella typhimurium"
FT /strain="LT2"
FT /map="92 minutes"
FT CDS complement(<1..312)
FT /codon_start=1
FT /db_xref="SPTREMBL:085309"
FT /transl_table=11
FT /gene="uvrA"
FT /product="excision nuclease UvrA"
FT /protein_id="AAC26637.1"
FT /translation="MDKIEVRGARTHNLKNINFVIPRODKLIVVTGLSGSGKSSLAFDTL
FT YAEQRRYVESLSAYARQFLSLMEKPDVDHIEGLSPAISIEQKSTSHNPRSTVGTITEI
FT "
FT CDS 583..1083
FT /codon_start=1
FT /db_xref="SPTREMBL:085310"
FT /transl_table=11
FT /gene="ssb"
FT /product="single-strand binding protein"
FT /protein_id="AAC26638.1"
FT /translation="MILVGNPGQDPEVRYMPSGGAVANLTATSESWRDKQTGEMKEQT
FT EWHRRVMFGLAEVAGEYLRLLSSQVYIEGQLRTRKRTDQNCQERYTTTELTSADRRVMQI
FT LGGPKGGGAPAGGHNRLGSPQQPQQPQGGNQFNGGAQSRPQQSAPAPSKEPMDFDDO
FT IPF"
FT tRNA 1898..1970
FT /note="putative"
FT /product="tRNA-Thr"
XX
SQ Sequence 27290 BP; 7965 A; 5530 C; 6661 G; 7134 T; 0 other;
gatctcggta atagtaccga ccgtagagcg cgggttgtgc gatgtcgatt tctgttcaat 60
tgagatcgcg ggcgatagcc cctcaatatg gtcgacatcc ggtttttcca tgagcgacaa 120
aaactgccgc gcgtaagcgg agagcgattc aacgtaacga cgctgccctt cggcatacag 180
ttgttattac gtatcatggc aggcctgagaa gcttttcaga agaggacact tataaaataa 2520
aggcttggtt agaagacaaa atcaatagta atttattgat agaaatggtt attcctcagg 2580
//
```

EMBL Header

Annotation

Sequence

Artemis

- Sequence viewer and analysis tool
 - Visualization of sequence features
 - DNA
 - Six frame translation
 - Perform and view analysis
 - Basic analysis
 - Launch more complex analysis and searches
 - Import and view the results of other searches

Outline of Artemis demonstration

- Artemis window features
- Open a genome sequence
- Changing the view
- Getting around
 - Goto Menu
 - Navigator
 - Feature Selector
- Basic analysis
 - Edit a feature
 - Fasta search
 - Show feature plots



Artemis

Drop Down Menus
Entry Button Line

Main Sequence
View Panel

Magnified
Sequence View
Panel

Feature Menu

Artemis Entry Edit: S_typhi.dna

File Entries Select View Goto Edit Create Write Run Graph Display (Noddy)

Selected feature: bases 930 amino acids 309 STY0003 (/class="3.1.18" /colour=7 /ec orthologue=K

Entry: ☐ S_typhi.dna ☒ S_typhi.tab

STY0002

STY0003 STY0004

misc_feature 800 1600 misc_feature 2400 3200 misc_feature 4000 4800 misc_feature 5600 6400 misc_feature 7200

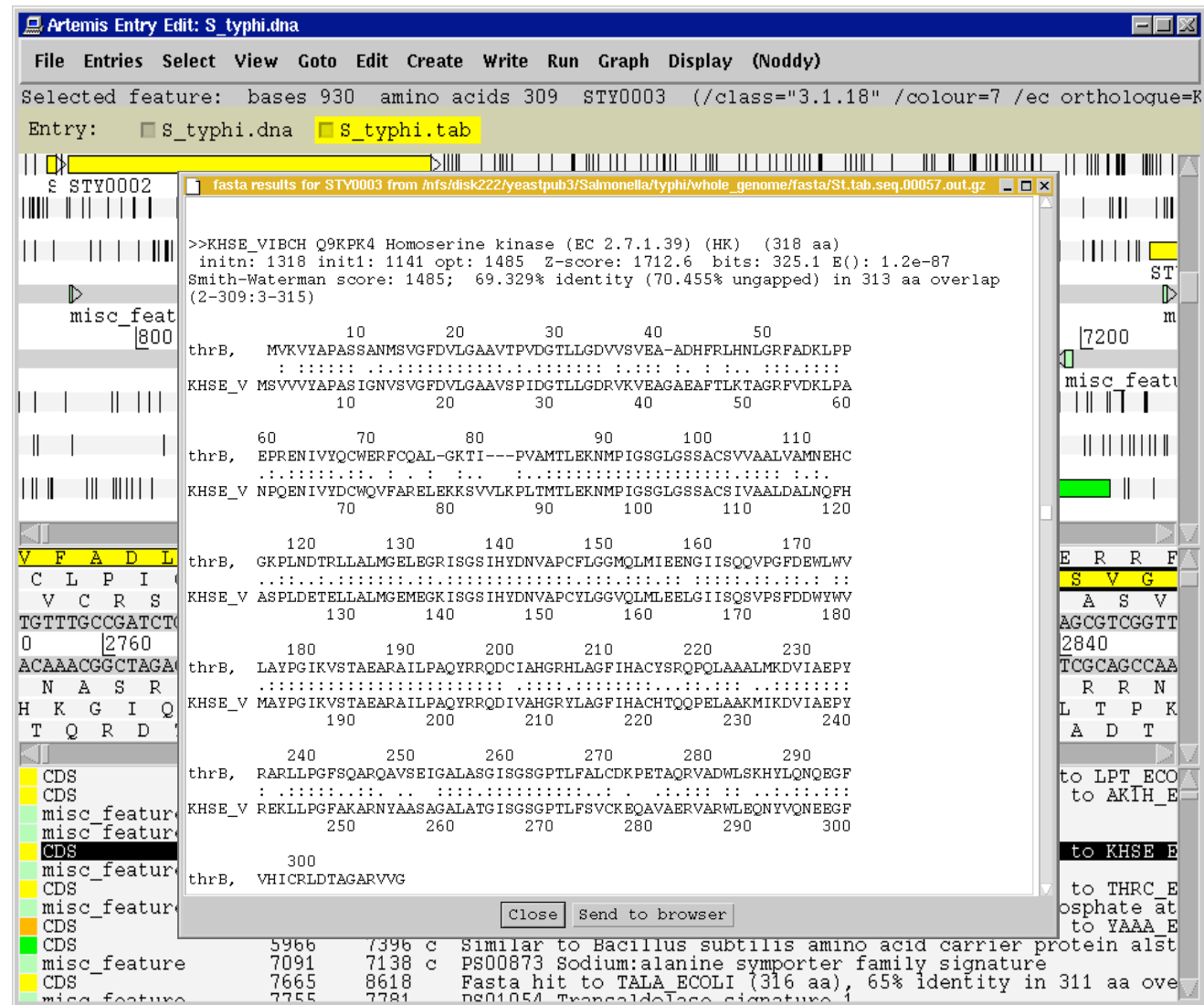
STY0005 STY0006

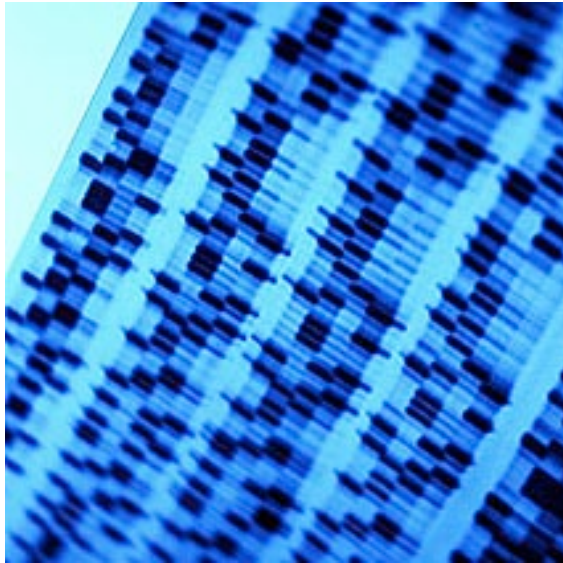
V F A D L L R T L S W K L G V # H G E S V C P G F Q R E H E R R F
C L P I C Y G P S H G S + E F N M V K V Y A P A S S A N M S V G
V C R S V T D P L M E V R S L T W * K C M P R L P A R T * A S V
TGTTGCCGATCTGTTACGGACCCTCTCATGGAAGTTAGGAGTTTAACATGGTGAAGTGTATGCCCCGGCTTCCAGCGCGAACATGAGCCTCGGTT
0 2760 2770 2780 2790 2800 2810 2820 2830 2840
ACAAACGGCTAGACAATGCCTGGGAGAGTACCTTCAATCCTCAAATTGTACCACTTTACATACGGGGCCGAAGGTCGCGCTTGTTACTCGCAGCCAA
N A S R N R V R E H F N P T # C P S L T H G P K W R S C S R R N
H K G I Q # P G E * P L # S N L M T F T Y A G A E L A F M L T P K
T Q R D T V S G R M S T L L K V H H F H I G R S G A R V H A D T

CDS	190	255	Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECOLI
CDS	337	2799	Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit to AK1H_ECOLI
misc_feature	343	369	PS00324 Aspartokinase signature
misc_feature	2314	2382	PS01042 Homoserine dehydrogenase signature
CDS	2801	3730	Orthologue of E. coli thrB (KHSE_ECOLI); Fasta hit to KHSE_ECOLI
misc_feature	3068	3103	PS00627 GHMP kinases putative ATP-binding domain
CDS	3734	5020	Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_ECOLI
misc_feature	4022	4066	PS00165 Serine/threonine dehydratases pyridoxal-phosphate at
CDS	5114	5887	Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI
CDS	5966	7396	Similar to Bacillus subtilis amino acid carrier protein alst
misc_feature	7091	7138	PS00873 Sodium:alanine symporter family signature
CDS	7665	8618	Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa ove
misc_feature	7755	7781	PS01054 Transaldolase signature 1

Sliders

Sliders





1. Gene prediction
2. Overview and pitfalls

Gene prediction programs: ORFs and CDSs

ORFs are not equivalent to CDSs

Not all open reading frames are coding sequences

Artemis demonstration with *Campylobacter jejuni* and *Streptomyces coelicolor*



Gene prediction strategy

- Gene Prediction programs
 - Orpheus
 - Glimmer

Human refinement

- Plots
 - Codon usage
 - Correlation score
 - GC plot
- tBlastx



Gene prediction programs

- Gene Prediction programs
 - Orpheus – Hidden Markov model
 - Glimmer – Interpolated Markov model
- Training set
 - tBlastx
 - Long ORF
 - Related genome gene set
- Looks for regions with properties similar to training set



Gene prediction programs: Statistics

Organism	Size (Mb)	G+C	CDS prediction				Final
			Glimmer ¹	G2 ¹	ORPHEUS ²	other	
<i>Campylobacter jejuni</i>	1.641	30.55	1761		1518	1783 ³	1654
<i>Neisseria meningitidis</i> A	2.184	51.81	3134		2024		2121
<i>Mycobacterium leprae</i>	3.268	57.80	949	5679 ⁴	4427		1605 intact 1115 pseudo
<i>Salmonella typhi</i>	4.809	52.09	5194		4666	4973 ⁵	4600
<i>Yersinia pestis</i>	4.654	47.64		2654	4312		4011

¹ <http://www.tigr.org/softlab/glimmer/glimmer.html>

² <http://pedant.mips.biochem.mpg.de/orpheus/index.html>

³ Start-to-stop >100 aa

⁴ TIGR CMR (<http://www.tigr.org/>)

⁵ GeneFinder (Krogh+Larson *pers comm*)



Gene prediction programs: Problems

ORFs are not equivalent to CDSs

Gene prediction programs find new genes that share properties with a given set of genes. They can be confounded by:

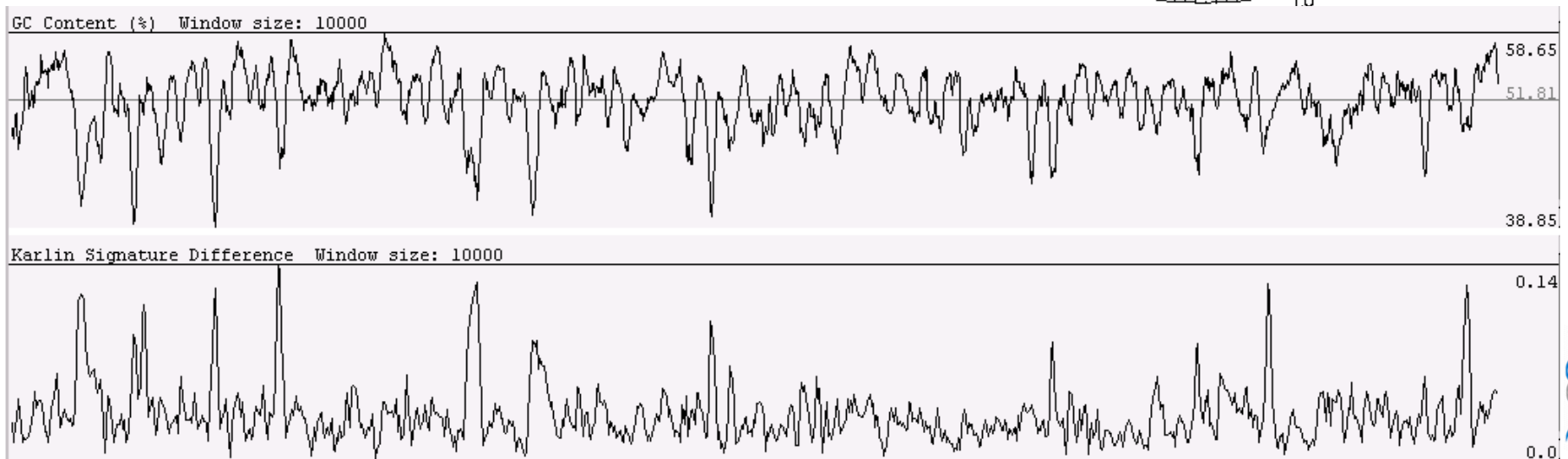
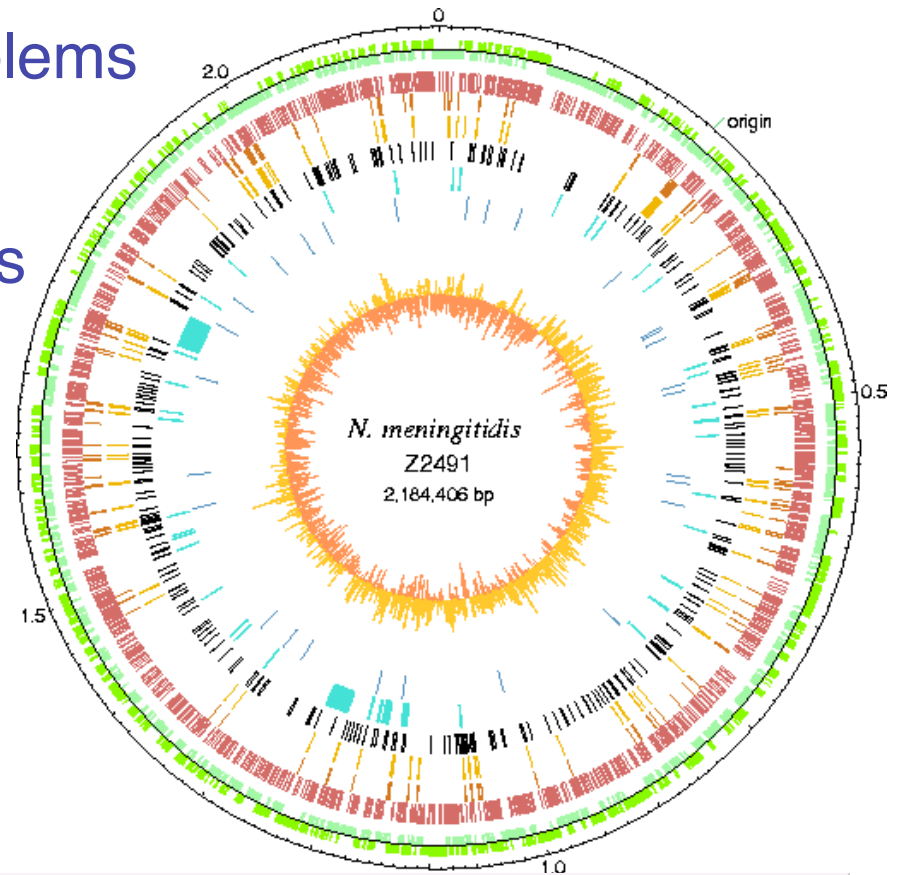
- Sequence constraints (ribosomal proteins etc.)
- Sequence biases
- Different sets of genes
- Horizontal gene transfer
- Non-coding DNA



Gene prediction programs: Problems

Sequence composition variation

N. meningitidis ribosomal proteins



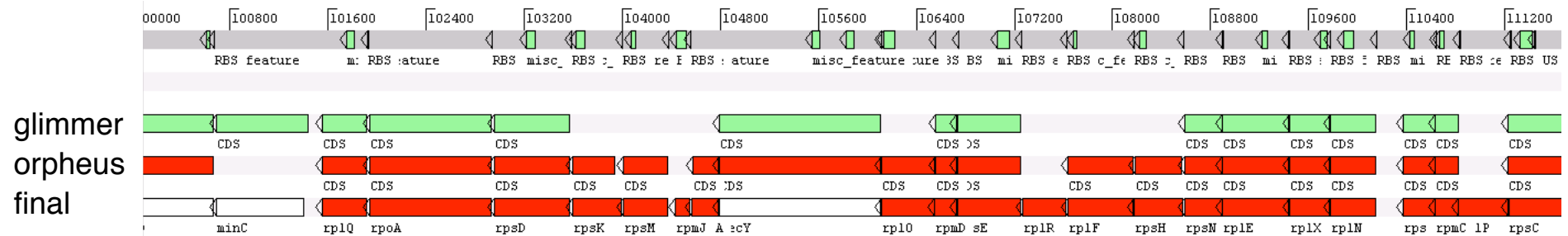
N. meningitidis ribosomal proteins



Gene prediction programs: Problems

Sequence composition variation

N. meningitidis ribosomal proteins

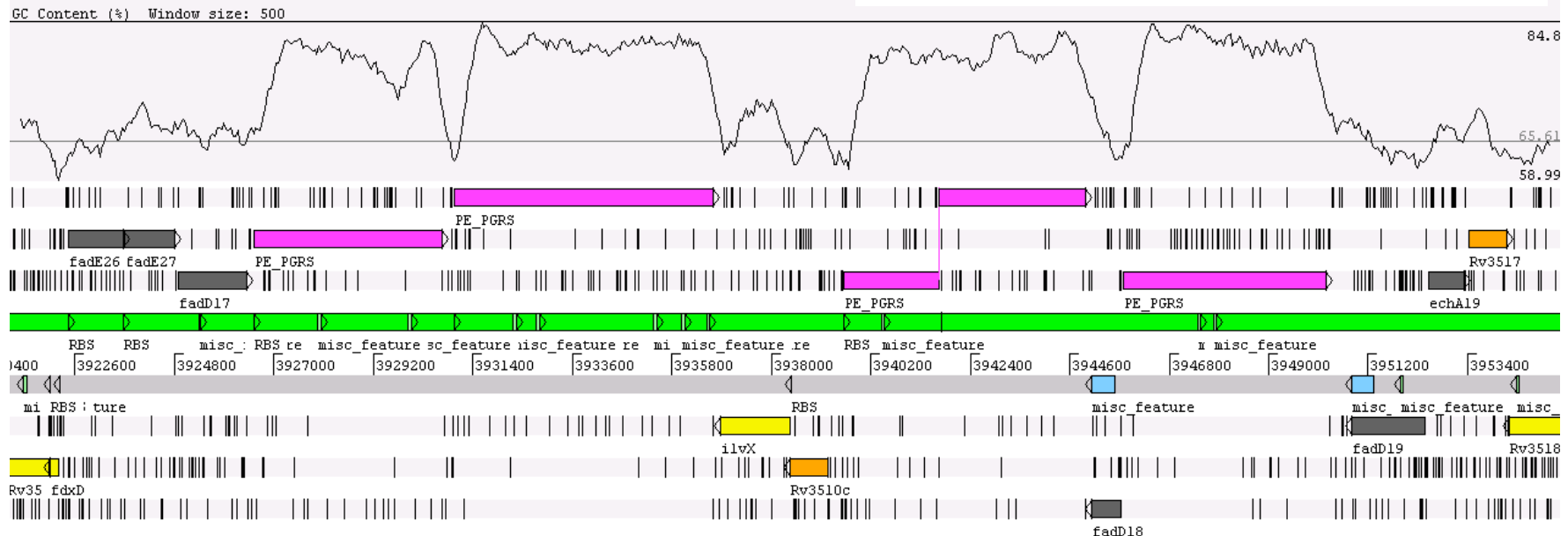
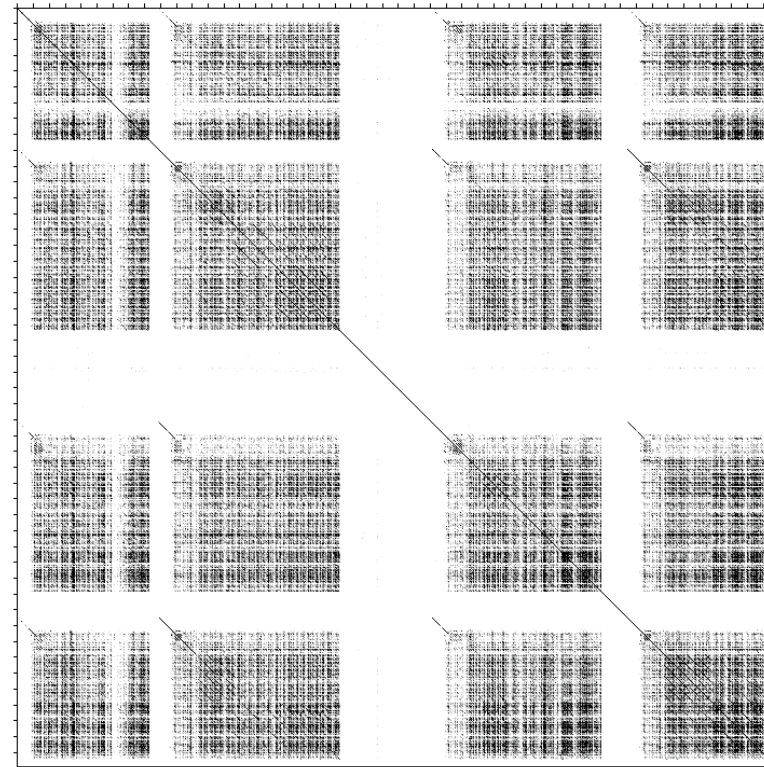


Gene prediction programs: Problems

Sequence composition variation

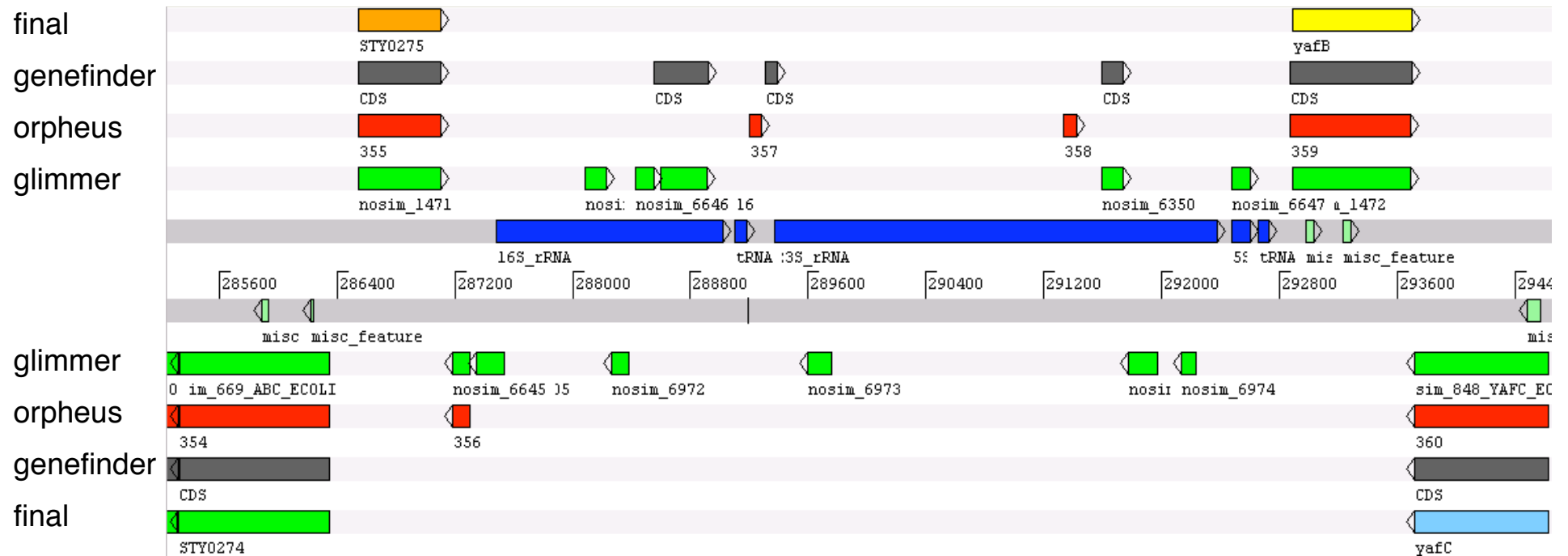
M.tuberculosis PGRS genes

Large, highly repetitive G+C rich genes.
Not predicted as coding by Markov-based programs



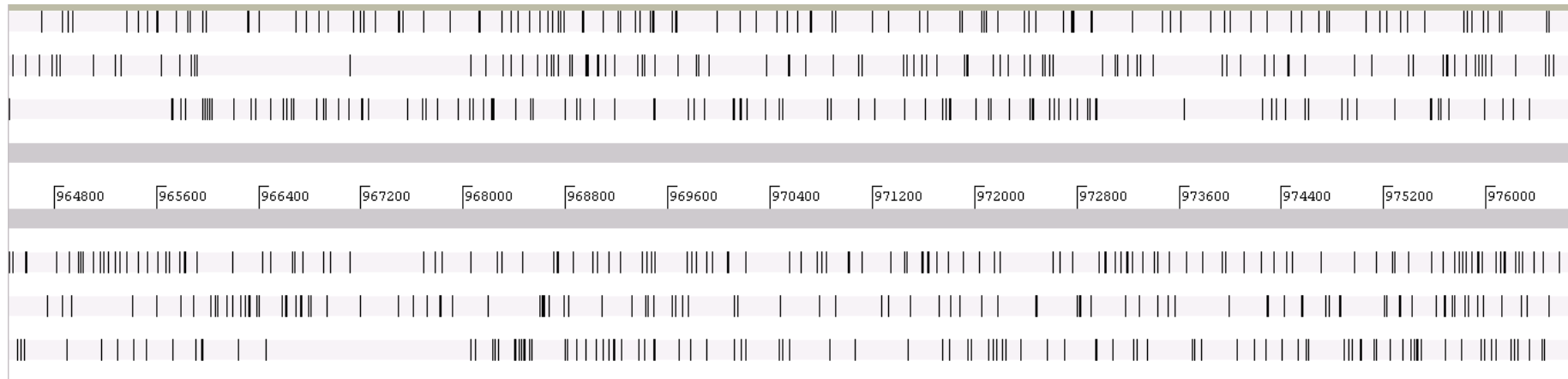
Gene prediction programs: Problems

Non-protein coding regions: *S. typhi* ribosomal RNA genes



Gene prediction programs: Problems

Pseudogenes: *M. leprae*

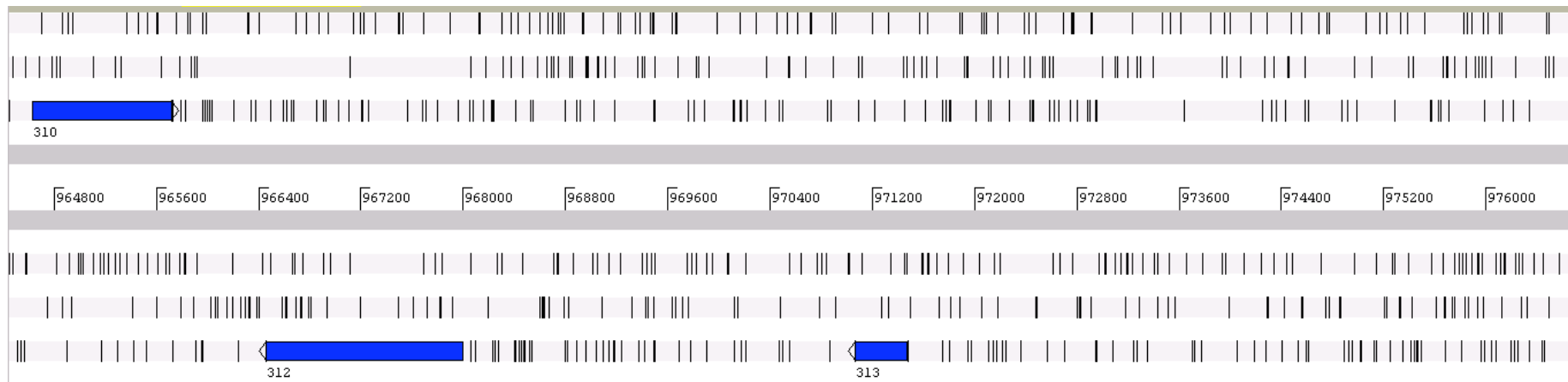


Gene prediction programs: Problems

Pseudogenes: *M. leprae*

Glimme

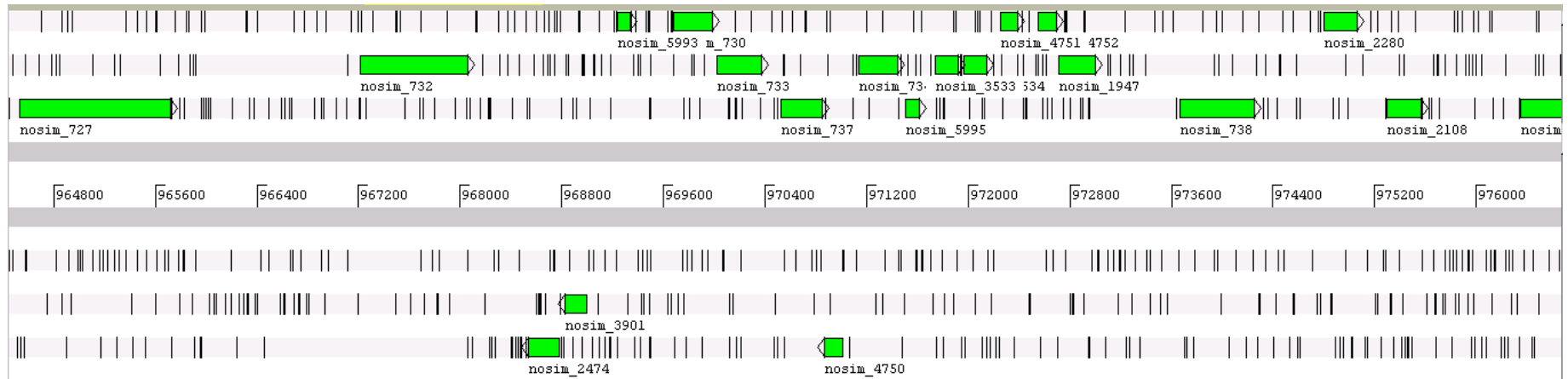
r



Gene prediction programs: Problems

Pseudogenes: *M. leprae*

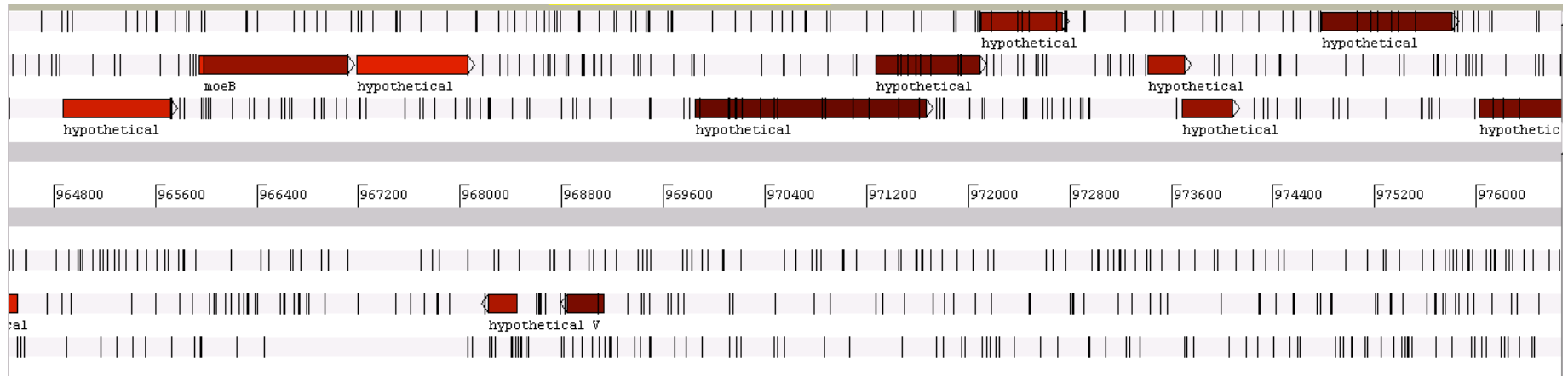
ORPHEUS



Gene prediction programs: Problems

Pseudogenes: *M. leprae*

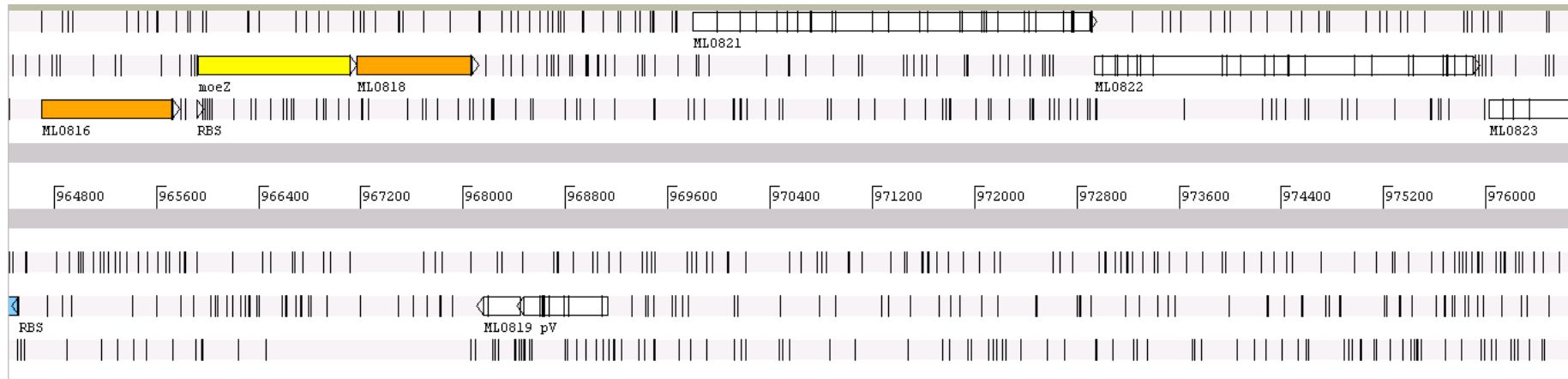
WUBLASTX vs. *M. tuberculosis*



Gene prediction programs: Problems

Pseudogenes: *M. leprae*

Final annotation



Prokaryotic gene prediction strategy

- Gene Prediction programs
 - Orpheus
 - Glimmer
- Plots
 - Codon usage
- tBlastx



Codon Usage tables

UUU 9.0 (3967)	UCU 5.4 (2392)	UAU 8.1 (3574)	UGU 4.1 (1816)
UUC 20.0 (8826)	UCC 11.6 (5114)	UAC 13.9 (6137)	UGC 7.5 (3298)
UUA 5.4 (2383)	UCA 6.9 (3057)	UAA 0.9 (402)	UGA 1.7 (744)
UUG 23.4 (10300)	UCG 18.5 (8133)	UAG 1.1 (464)	UGG 13.7 (6045)
CUU 8.8 (3883)	CCU 7.3 (3209)	CAU 8.6 (3800)	CGU 12.9 (5673)
CUC 14.5 (6367)	CCC 13.0 (5734)	CAC 14.8 (6505)	CGC 22.1 (9752)
CUA 8.8 (3873)	CCA 9.6 (4225)	CAA 11.3 (4992)	CGA 9.6 (4230)
CUG 37.8 (16646)	CCG 23.9 (10512)	CAG 22.1 (9715)	CGG 19.9 (8755)
AUU 11.8 (5177)	ACU 10.8 (4742)	AAU 8.2 (3618)	AGU 6.7 (2952)
AUC 30.8 (13565)	ACC 28.7 (12627)	AAC 18.4 (8091)	AGC 13.6 (5989)
AUA 5.0 (2214)	ACA 8.4 (3716)	AAA 9.4 (4128)	AGA 2.7 (1190)
AUG 19.7 (8677)	ACG 14.3 (6301)	AAG 17.8 (7838)	AGG 4.1 (1810)
GUU 14.5 (6400)	GCU 21.2 (9314)	GAU 21.4 (9433)	GGU 23.3 (10265)
GUC 28.7 (12657)	GCC 39.8 (17526)	GAC 35.9 (15797)	GGC 33.2 (14603)
GUA 9.8 (4322)	GCA 17.9 (7888)	GAA 21.2 (9335)	GGA 11.8 (5195)
GUG 37.2 (16399)	GCG 34.4 (15164)	GAG 28.2 (12437)	GGG 14.5 (6383)

availble from <http://www.kazusa.or.jp/codon/>



Codon usage

