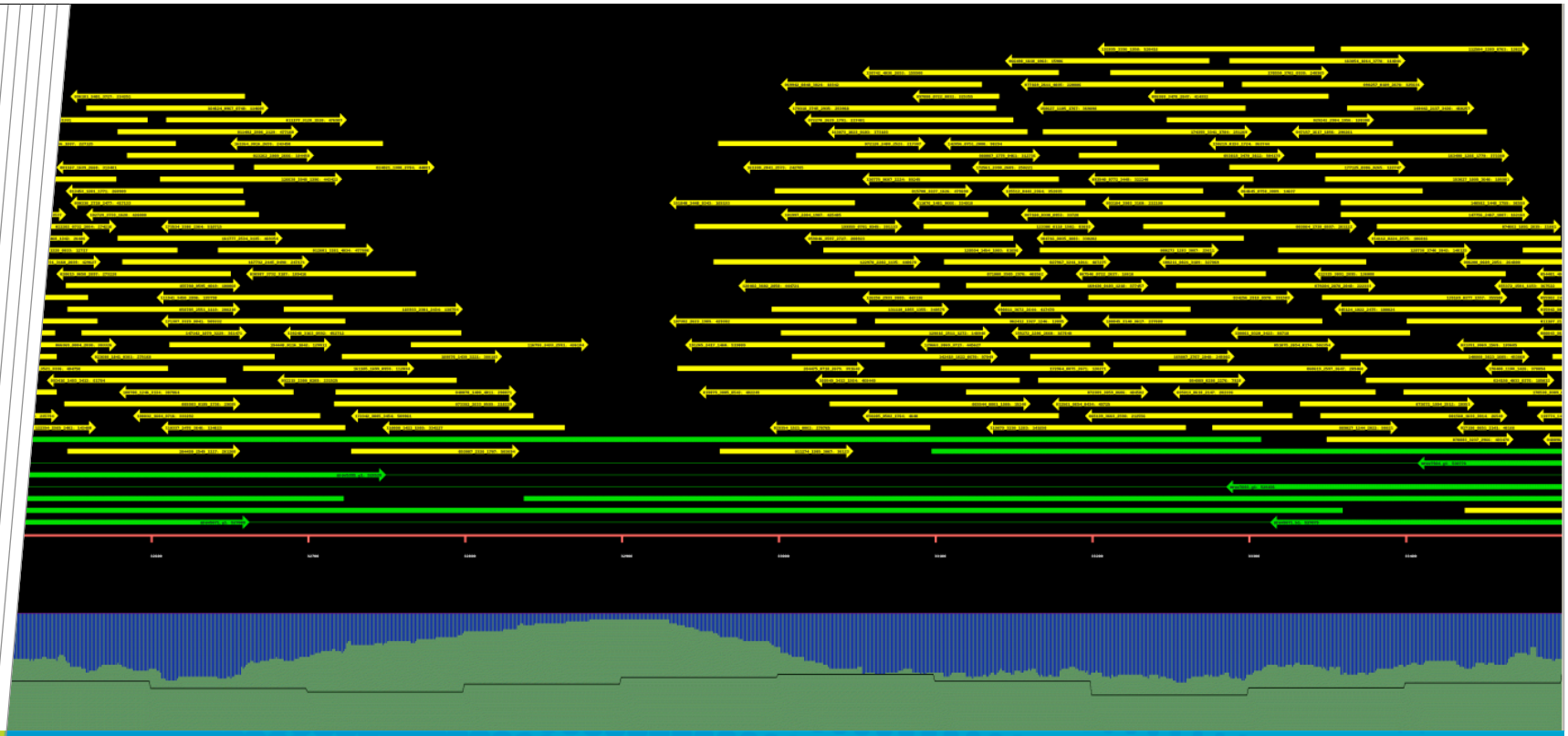


www.csiro.au

Querying Genomes

Using databases to answer biological questions

Paul Greenfield
CMIS Transformational Biology
28/10/2009

CSIRO

Databases Are Data Stores…

- Databases are widely used in genomics
  - Primarily as data stores
  - Retrieve strings into (big) in-memory data structures
  - Perform pattern-matching searches
    - Look for approximate matches for long-ish DNA strings
    - Look for gene patterns
  - Typically have simple retrieval queries against the database, complex external string-matching code
- Methods developed for a different world
  - Genomic data was scarce & expensive
- Genomic data is now plentiful and cheap(er)
  - Thousands of human genomes being sequenced
  - Sequencing whole microbial populations (metagenomics)

More Than Data Stores… An Experiment

- Answering biological questions with SQL queries?
  - What do these organisms have in common?
  - What organisms are living in this environment?
  - Inspired by Jim Gray's work on astronomy databases
- Improve speed of common tasks?
  - Query optimisation, parallelism, buffering, read-ahead, …
  - Need data base-friendly algorithms or genetics-friendly databases or both
- Currently working with bacteria
  - Small genomes (~5Mbp), little 'junk' or repetition
  - ~1000 sequenced, annotated species/strains
  - And using '# of shared 25-mers' as a comparison metric

A k-mer Bacterial Database…

- Species, Sequences, Genes
  - Metadata… start, end, length, strand, product, name, …
- All (tiled) 25-mers
  - Every possible 25-mer →
  - Species, sequence, location, mer
  - Indexed by mer (non-unique index)

ATCGAATTTCGTAATCGTACATGTTTAACCGTA…
ATCGAATTTCGTAATCGTACATGTT
TCGAATTTCGTAATCGTACATGTTT
CGAATTTCGTAATCGTACATGTTTA
GAATTTCGTAATCGTACATGTTTAA
AATTTCGTAATCGTACATGTTTAAC
ATTTCGTAATCGTACATGTTTAACC
TTTCGTAATCGTACATGTTTAACCG
TTCGTAATCGTACATGTTTAACCGT
TCGTAATCGTACATGTTTAACCGTA

- Counts of shared k-mers
  - Result of cross-matching k-mers from all bacteria
  - Genome to genome, genes
  - Gene to gene
- About 300GB for 720 species/strains
  - With all DNA sequences as uncompressed ATCG strings…

Why ‘shared 25-mers’?

- K-Mer space is big (for reasonable ‘k’)
  - 10^{15} 25-mers; 10^{30} 50-mers; even 10^{12} 20-mers
- Accidental collisions are rare in 25-mer space
  - ‘Unrelated’ organisms share few/none 25-mers
    - Apart from some highly conserved fundamental genes/RNA
  - And not just for bacteria…
- Shared 25-mers then…
  - … say something interesting about relationships, conservation or gene-swapping or …
    - Closely related organisms share many 25-mers
    - Count of shared 25-mers drops rapidly with taxonomic distance
    - … a relationship metric that can be used in queries

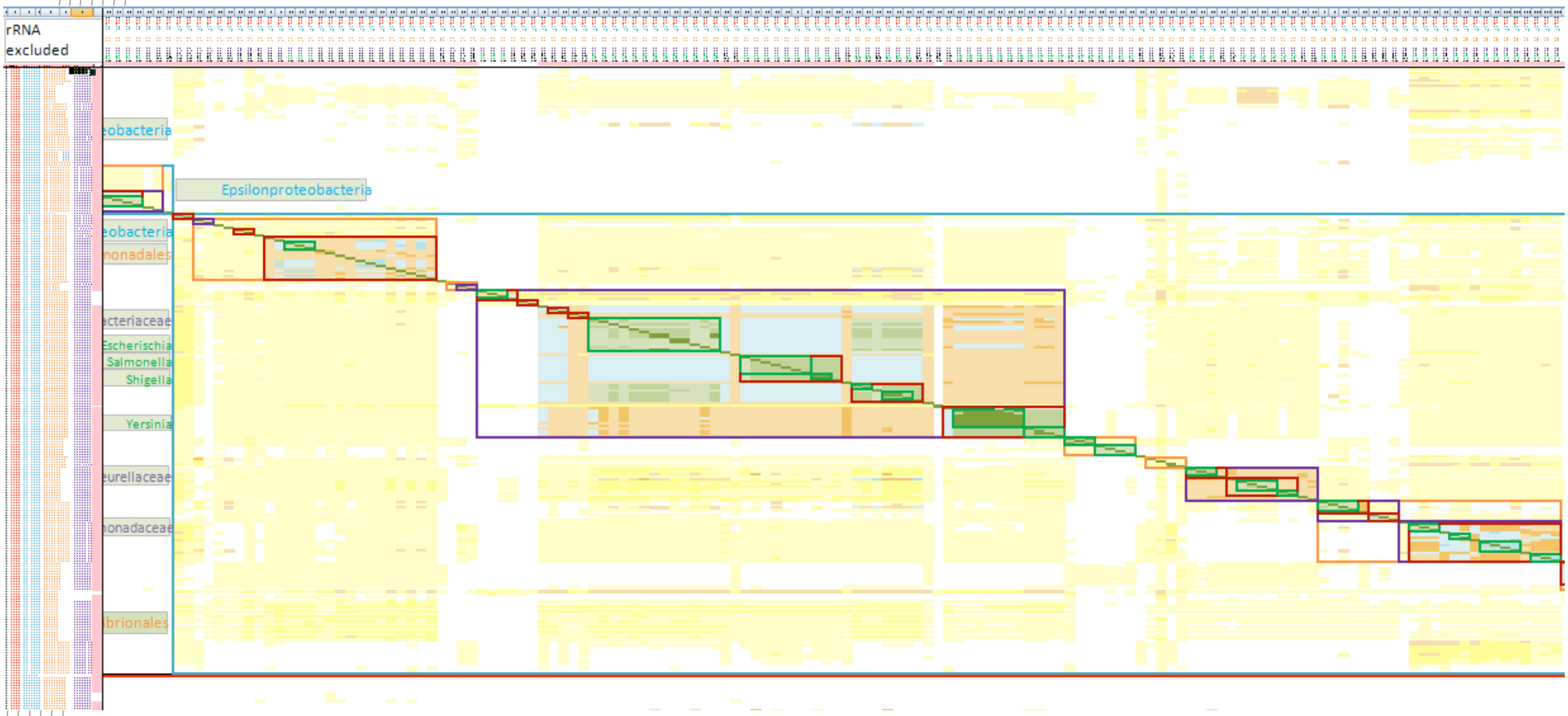
And… Exact k-mer Matching Is Fast

- Exact k-mer matching by…
  - Queries on indexed database tables
    - 100,000 lookups/second (for dense searches)
    - Need to structure DB & lookup code very carefully
    - Sparse searches perform less well (fewer buffer hits, random IO)
  - In-memory hash tables (quick hack for now)
    - Cache parts of database in app (sequential index read)
    - 500,000 lookups/second (and improving)
    - Dense & sparse searches perform at sequential read speed
- Fast enough for current purposes
  - Cross-compare all bacteria
    - All k-mers in all sequences compared to all k-mers in all sequences
    - Gather comparison counts to support queries…
  - Compare metagenomic sample to all bacteria
    - What species are present? Partition reads by species, …

Comparing All Bacteria

- Cross-compare all 25-mers from all bacteria
  - Generate (overlapping) k-mer ‘tiles’ from each bacteria
  - Compare these to the genomes of all other organisms
  - Count various interesting things when we find a match
- Practicalities
  - About 2.5 billion bases in available sequences
    - Double this because matches could be on other strand
  - 2.5 billion source k-mers & 5 billion target k-mers
- Actualities
  - Use pre-computed indexes, sorting, partitioning, hash caches
  - Index looks up against all target genomes at once
  - High-level all-to-all comparison in ~3 hours on workstation
  - ‘Dense’ search - all k-mers matched at least once

The View From 10,000m (25-mers)



- Taxonomic groups show up strongly
- Considerable ‘white space’ (few shared 25-mers)
- Distinct ‘shadows’ of other taxonomic groups

What Do The Faint Shadows Mean?

- Accidental collisions or genes?
  - Answered by looking at gene-level intersections
  - Count matches by tracking [source gene, target gene] hits

Deinococcus geothermalis DSM 11300 (1, 200) cobaltochelatase (CobN subunit) 4335bp onto…

	0.20					
7	0.30 %	Mycobacterium avium 104	1	4118	cobaltochelatase	3576
9	0.20 %	Mycobacterium avium paratuberculosis	1	3197	cobaltochelatase	3576
8	0.20 %	Mycobacterium bovis BCG Pasteur	1	3660	cobaltochelatase	3585
8	0.20 %	Mycobacterium bovis 1173P2	1	3645	cobaltochelatase	3585
8	0.20 %	Mycobacterium smegmatis MC2 155	1	6524	cobaltochelatase	3636
8	0.20 %	Mycobacterium tuberculosis CDC1551	1	3785	cobaltochelatase	3588
8	0.20 %	Mycobacterium tuberculosis F11	1	3683	cobaltochelatase	3585
8	0.20 %	Mycobacterium tuberculosis H37Ra	1	3739	cobaltochelatase	3585
8	0.10 %	Mycobacterium tuberculosis H37Rv	1	3697	cobaltochelatase	3585
2	0.40 %	Mycobacterium vanbaalenii PYR-1	1	5831	cobaltochelatase	3618
14	0.10 %	Nocardia farcinica IFM10152	1	5583	cobaltochelatase	3627
2	0.40 %	Saccharopolyspora erythraea NRRL 2338	1	10672	cobaltochelatase	3606
15	0.40 %	Streptomyces avermitilis	1	11778	cobaltochelatase	3654
16	%	Streptomyces coelicolor	1	3324	cobaltochelatase	3654

What's in the faintest shadows?

Campylobacter hominis ATCC BAA-381\NC 009714 (1711125) compared to Helicobacter pylori Shi470 (0.01087% & 0.0109%)

Gene	geneID	str	len	product	Gene	geneID	str	len	product	+	RC
96	non-coding	1	4408	rRNA	1754	non-coding	1	4464	rRNA	0	482
96	non-coding	1	4408	rRNA	2393	non-coding	1	4581	rRNA	0	482
121	non-coding	1	3659	rRNA	1869	non-coding	1	529	non-coding	55	0
121	non-coding	1	3659	rRNA	1967	non-coding	1	2214	rRNA	0	315
…	non-coding			…					…		
710	non-coding	1	5618	rRNA	1869	non-coding	1	529	non-coding	55	0
…	non-coding			…		…			…		
712	non-coding	1	4520	rRNA	2393	non-coding	1	4581	rRNA	0	482
756	154149469	1	1932	putative cell division protease FtsH-like protein	658	188527188	-1	1899	cell division protein FtsH	0	2
1157	154148614	1	1023	Holliday junction DNA helicase B	673	188527198	-1	1011	Holliday junction DNA helicase B	0	4
1199	non-coding	1	382	non-coding	400	non-coding	1	239	non-coding	20	0
1226	154148152	1	822	enoyl-(acyl carrier protein) reductase	332	188527000	1	828	enoyl-(acyl carrier protein) reductase	1	0
1313	154147926	-1	1059	elongation factor Ts	2658	188528329	-1	1068	elongation factor Ts	2	0
1439	non-coding	1	127	non-coding	1674	non-coding	1	411	non-coding	1	0
2107	154149028	1	1296	transcription term. factor Rho	1393	188527605	1	1317	transcription term. factor Rho	5	0
2220	non-coding	1	1327	non-coding	400	non-coding	1	239	non-coding	20	0
2220	non-coding	1	1327	non-coding	1777	non-coding	1	919	non-coding	13	0
2367	non-coding	1	3207	rRNA	1869	non-coding	1	529	non-coding	0	55
2367	non-coding	1	3207	rRNA	2458	non-coding	1	2062	rRNA	315	0
2449	non-coding	1	508	non-coding	2042	non-coding	1	326	non-coding	9	0
2492	non-coding	1	1829	non-coding	11	non-coding	1	3200	non-coding	6	0
2593	non-coding	1	1573	non-coding	1674	non-coding	1	411	non-coding	1	0
2593	non-coding	1	1573	non-coding	1676	non-coding	1	2349	non-coding	5	0
2624	non-coding	1	725	non-coding	2090	non-coding	1	722	non-coding	56	0
2758	154148859	-1	1809	GTP-binding protein TypA/BipA	786	188527265	1	1800	hypothetical protein	0	4

Justifying the ‘shared k-mers’ Metric

- Most of the BxB matrix is uncoloured
  - Less than 0.001% shared 25-mers
- Shadows of various hues (higher sharing)
  - Only significant sharing for close relatives
  - Sharing highly non-random, reflecting shared genes
  - Genes with shared k-mers normally have same annotations
- ‘Accidental’ k-mer collisions are rare
  - Shared k-mers seem to almost always mean something
  - Even when only a very few are shared…
  - Highly-conserved parts of critical genes?
  - Recent horizontal gene transfers??

Where Did These Tables Come From?

- Results of single SQL queries
  - Run all-bacteria-to-all-bacteria comparison
  - Track source (gene) and target (gene) for each match
  - Load intersections into database as genome-onto-gene & gene-onto-gene hits tables
  - Query to get gene-level matches between organisms
- Answering biological questions with queries
  - What genes are shared across families?
  - How can we distinguish Citrobacter from its relatives?
  - Do the shadows represent shared genes?
  - Can we find functions for hypothetical genes?
  - What genes are unique to an organism
  - ….

Finding Conserved Genes

```
-- find most conserved genes
select s.SpeciesID as 'Source Species', g.SequenceNo, t.SpeciesID as 'Target Species',
        gh.targetGeneNo, g.GeneID, gh.MerHits,  g.GeneStart, g.GeneEnd, g.GeneLength,
        convert(float, gh.MerHits)/convert(float,g.GeneLength) as '%hits',  a.Product
from GeneHits gh, Species s, Species t, Genes g, BacteriaAnnotations.dbo.Annotations a
where gh.sourceSpeciesNo=s.SpeciesNo AND s.SpeciesID in
(select b.SpeciesID from BacteriaAnnotations.dbo.SpeciesNames b where b.OrderID='Enterobacteriales' )
and gh.targetSpeciesNo=t.SpeciesNo and t.SpeciesID in
(select c.SpeciesID from BacteriaAnnotations.dbo.SpeciesNames c where c.GenusID='Citrobacter' )
and g.geneNo=gh.targetGeneNo and g.speciesNo=gh.targetSpeciesNo and g.SequenceNo=gh.targetSequenceNo
and convert(float, gh.MerHits)/convert(float,g.GeneLength) > 0.8 -- more than 80% shared 25-mers
and g.GeneID=a.PID
and g.GeneLength>=25
order by t.SpeciesID, g.SequenceNo, gh.targetGeneNo
```

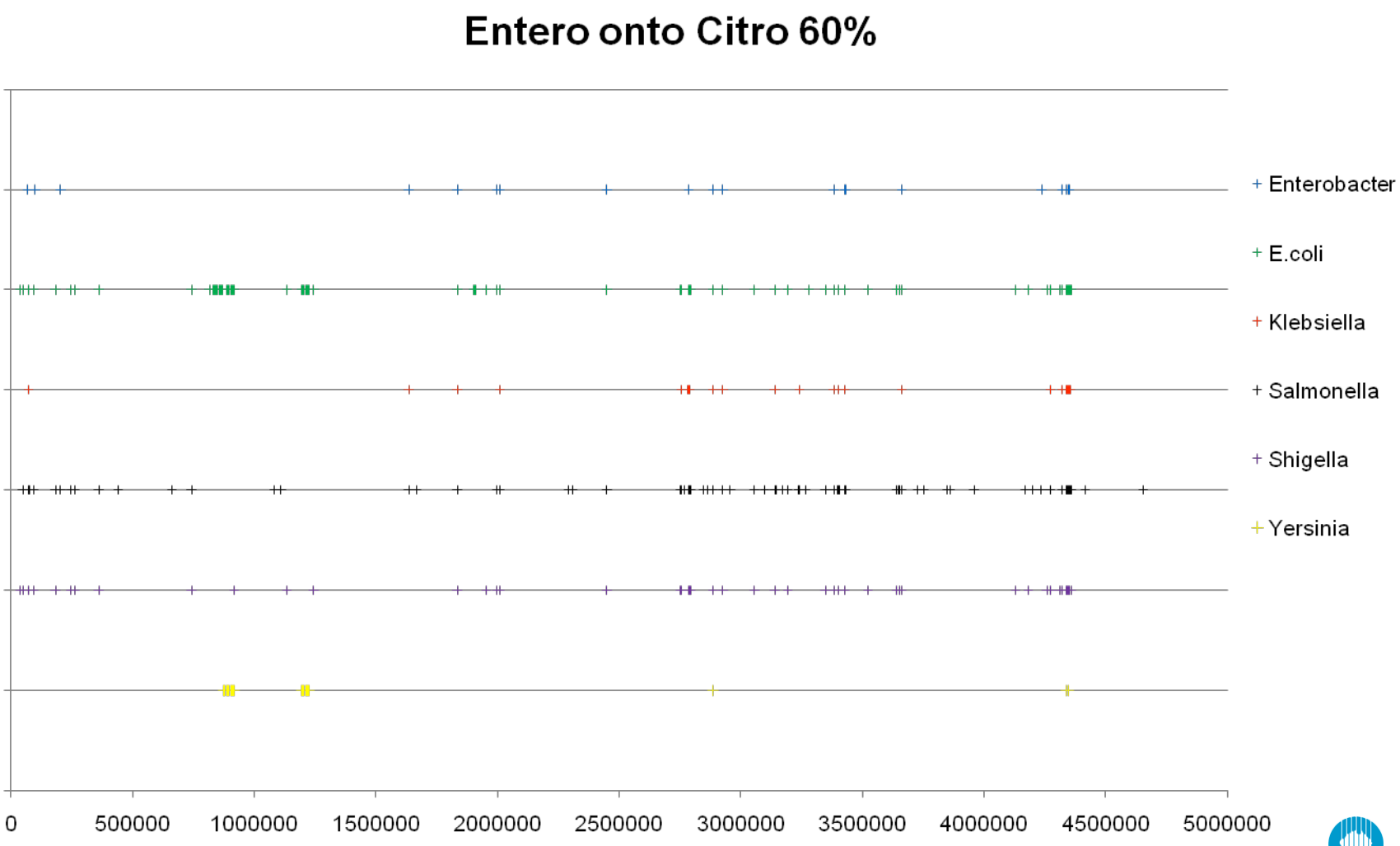
Genes from any of the Enterobacteriales that have matches with any Citrobacter at better than 80% coverage

What Is Highly Conserved?

- All Salmonella genes shared at > 90% with (Enterobacteriales excluding Salmonella)

49	hypothetical protein	1	22085	22822	738	92%	Escherichia_coli_SECEC_SMS_3_5	260	IS26 transposase	-1	124842	125564	723
49	hypothetical protein	1	22085	22822	738	93%	Escherichia_coli_SECEC_SMS_3_5	264	IS26 transposase	1	127034	127756	723
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	22	transposase for IS26	-1	9276	9992	717
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	6	putative transposase	-1	1008	1724	717
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	33	putative transposase	-1	10667	11383	717
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	53	hypothetical protein	-1	23121	23837	717
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	127	putative transposase	1	63307	64023	717
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	138	putative transposase	-1	67105	67821	717
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	147	putative transposase	1	71654	72370	717
49	hypothetical protein	1	22085	22822	738	92%	Klebsiella_pneumoniae_MGH_78578	173	putative transposase	-1	84446	85162	717
52	hypothetical protein	-1	23420	24433	1014	98%	Escherichia_coli_APEC_O1	152	IntI1 integrase	-1	95461	96474	1014
52	hypothetical protein	-1	23420	24433	1014	95%	Escherichia_coli_SECEC_SMS_3_5	198	integrase/recombinase	-1	90229	91242	1014

Where are these highly shared genes?



Finding Unique Genes

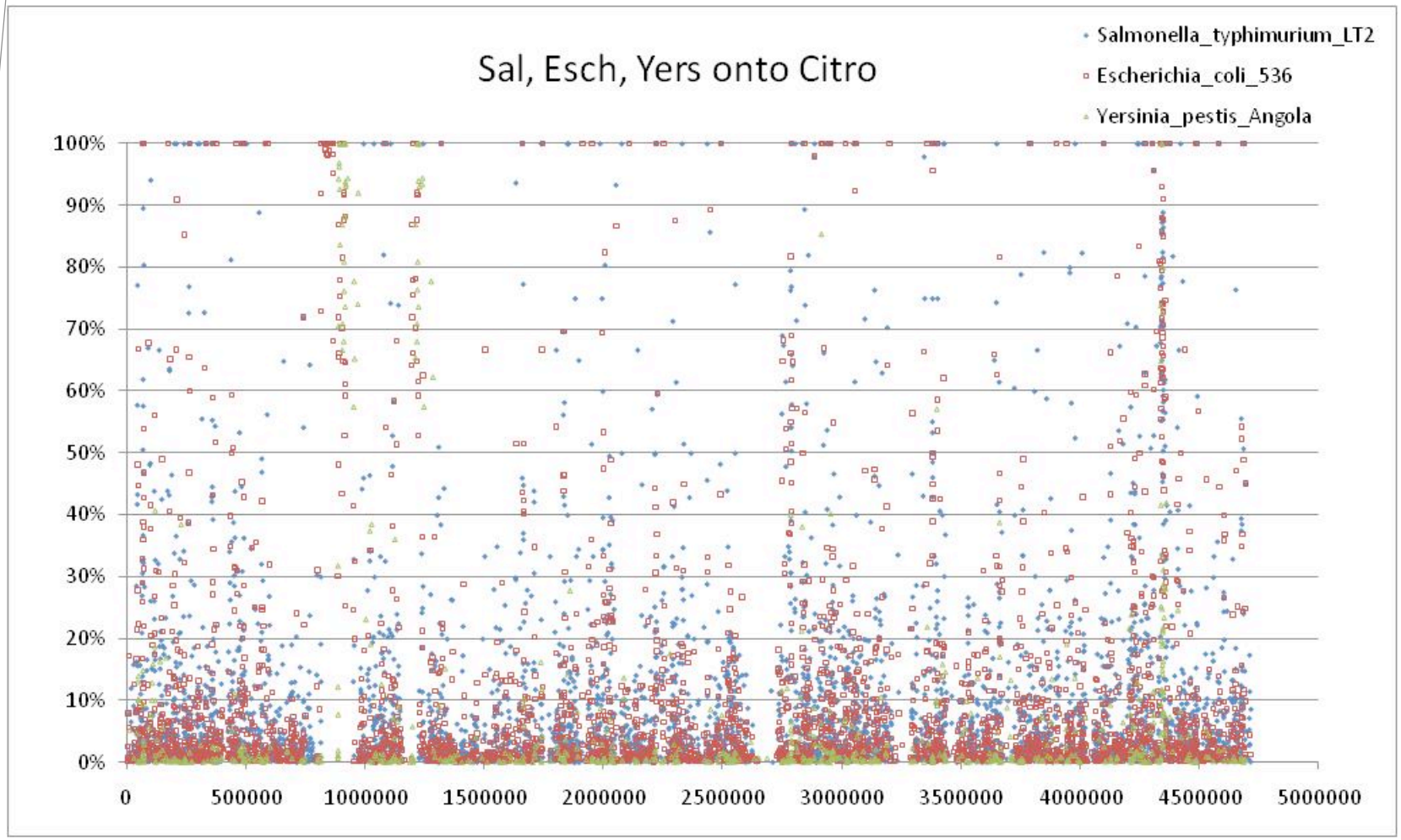
SQLQuery1.sql - not connected* | BacteriaDB qu...not connected*

```sql
-- find genes unique to an organism
select s.SpeciesNo, g.GeneNo, g.GeneID, g.GeneStart, g.GeneEnd, g.GeneLength, a.Product
from Species s, Genes g, BacteriaAnnotations.dbo.Annotations a
where g.GeneNo in (select gt.GeneNo         -- all genes in target organism (species/sequence relative gene no. set)
                   from Genes gt, Species st
                   where gt.SpeciesNo=st.SpeciesNo and st.SpeciesID='Citrobacter_koseri_ATCC_BAA-895' and gt.SequenceNo=1
                   except
                   select distinct gs.GeneNo            -- minus those genes that have sufficient hits from other organisms
                   from Genes gs, Species ss, GeneHits gh, Species st
                   where gh.sourceSpeciesNo=ss.SpeciesNo and -- start with gh rows from source organisms
                         ss.SpeciesID in (select b.SpeciesID from BacteriaAnnotations.dbo.SpeciesNames b where b.OrderID='Enterobacte
                         gh.targetSpeciesNo=st.SpeciesNo and -- limit to those mapping onto our target organism/sequence
                         st.SpeciesID='Citrobacter_koseri_ATCC_BAA-895' and gh.TargetSequenceNo=1 and
                         gs.geneNo=gh.targetGeneNo and gs.speciesNo=gh.targetSpeciesNo and gs.SequenceNo=gh.targetSequenceNo and --
                         convert(float, gh.MerHits)/convert(float,g.GeneLength) > 0.01   -- and %hit > 1%
                   ) and
      g.SpeciesNo=s.SpeciesNo and s.SpeciesID='Citrobacter_koseri_ATCC_BAA-895' and g.SequenceNo=1 and
      g.GeneID=a.PID and
      g.GeneID <> 'non-coding'          -- only 'genes'
order by g.GeneNo
```

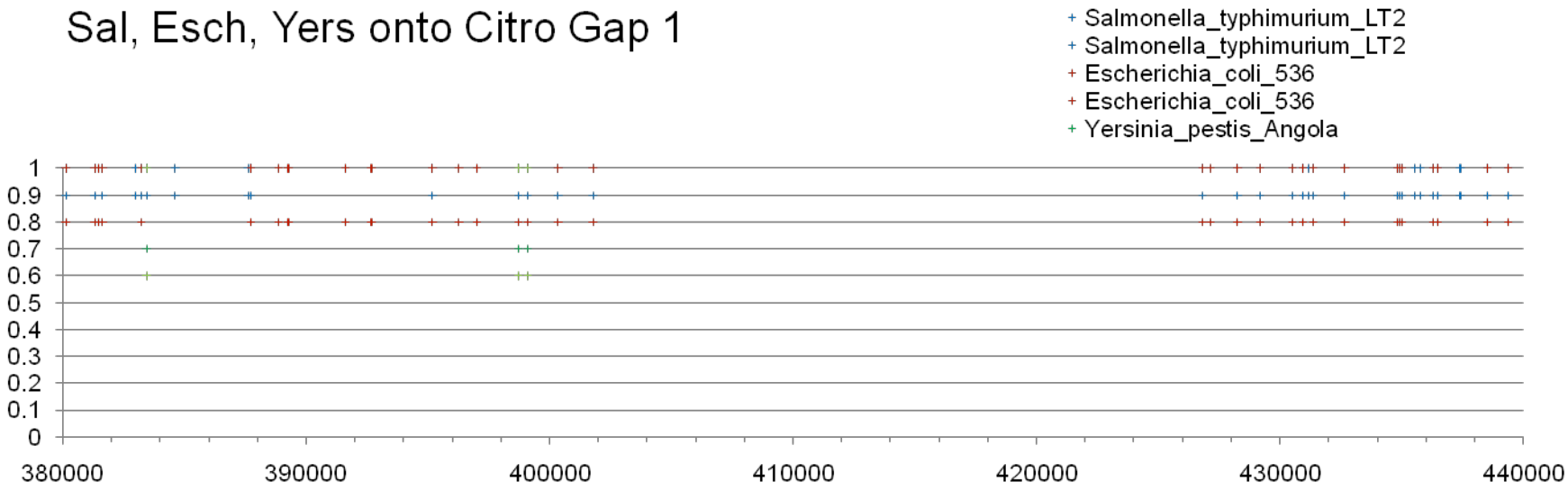
(All genes in Citrobacter)

- (those Citrobacter genes that have shared 25-mers with other members of the Enterobacteriales order at more than 1% coverage)

Distinguishing Citrobacter



Recognising Citrobacter



SpeciesID	SourceSeq	MerHits	MersHits NonrRNA	Target Seq	Target GeneNo	Gene Length	GeneStart	GeneEnd	gap	%hits	Product
Agrobacterium tumefaciens C58 Cereon	2	35	35	1	700	1896	412547	414442		1.8%	hypothetical protein
Bradyrhizobium BTAi1	2	8	8	1	700	1896	412547	414442		0.4%	hypothetical protein
Bradyrhizobium ORS278	1	11	11	1	700	1896	412547	414442		0.6%	hypothetical protein
Chromobacterium violaceum	1	2	2	1	700	1896	412547	414442		0.1%	hypothetical protein
Hahella chejuensis KCTC 2396	1	3	3	1	713	2544	419013	421556		0.1%	cAMP phosphodiesterase
Parvibaculum lavamentivorans DS-1	1	2	2	1	703	681	414783	415463		0.3%	hypothetical protein
Pseudomonas putida GB 1	1	5	5	1	700	1896	412547	414442		0.3%	hypothetical protein
Rhodopseudomonas palustris CGA009	1	5	5	1	700	1896	412547	414442		0.3%	hypothetical protein
Rhodopseudomonas palustris TIE 1	1	8	8	1	700	1896	412547	414442		0.4%	hypothetical protein
Thiomicrospira crunogena XCL-2	1	2	2	1	713	2544	419013	421556		0.1%	cAMP phosphodiesterase

Finding Matches For Hypothetical Genes

BacteriaDB qu...\gre403 (54))*

```
-- find shared genes where source is hypothetical but target is known
select s.SpeciesID as 'Source', gh.SourceSpeciesNo, gh.SourceSequenceNo, gh.SourceGeneNo,
       sg.GeneLength, sg.GeneStart, sg.GeneID, sa.Product,
       t.SpeciesID as 'Target', gh.TargetSpeciesNo, gh.TargetSequenceNo, gh.TargetGeneNo,
       tg.GeneLength, tg.GeneStart, tg.GeneID, ta.Product,
       gh.MerHits
from GeneToGeneHits gh, Genes sg, Genes tg, Species s, Species t,
     BacteriaAnnotations.dbo.Annotations sa, BacteriaAnnotations.dbo.Annotations ta
where gh.MerSize=25 and
      gh.sourceSpeciesNo=s.SpeciesNo AND s.SpeciesID = 'Lactococcus_lactis_cremoris_SK11' and
      sg.SpeciesNo=gh.SourceSpeciesNo and sg.SequenceNo=gh.SourceSequenceNo and sg.GeneNo=gh.SourceGeneNo and
      sg.GeneID=sa.PID and
      t.SpeciesNo=gh.TargetSpeciesNo and
      tg.SpeciesNo=gh.TargetSpeciesNo and tg.SequenceNo=gh.TargetSequenceNo and tg.GeneNo=gh.TargetGeneNo and
      tg.GeneID=ta.PID and
      sa.Product='hypothetical protein' and ta.Product<>'hypothetical protein'
order by s.SpeciesID, gh.SourceSequenceNo, sg.GeneStart, t.SpeciesID, gh.TargetSequenceNo, gh.TargetGeneNo
```

Results | Messages

	Target	TargetSpeciesNo	TargetSequenceNo	TargetGeneNo	GeneLength	GeneStart	GeneID	Product	MerH
1	Leuconostoc_citreum_KM20	335	2	3	522	1429	170016275	RepB-like protein	11
2	Rickettsia_prowazekii	534	1	306	667	203372	non-coding	non-coding	1
3	Sulfurihydrogenibium_YO3AOP1	647	1	1369	1773	861804	188996758	Highly conserved protein containing a thioredox...	2
4	Lactococcus_lactis_cremoris_SK11	321	4	43	3649	27540	non-coding	non-coding	6
5	Streptococcus_gordonii_Challis_substr_CH1	615	1	1616	174	962683	157150223	HsdD protein	4
6	Streptococcus_gordonii_Challis_substr_CH1	615	1	1617	228	962832	157150137	HsdD protein	112
7	Lactococcus_lactis_cremoris_MG1363	320	1	1194	185	656895	non-coding	non-coding	6
8	Oenococcus_oeni_PSU-1	437	1	585	321	346140	non-coding	non-coding	18
9	Streptococcus_thermophilus_CNRZ1066	638	1	1464	237	776000	55822801	dipeptidase. truncated	6
10	Streptococcus_thermophilus_CNRZ1066	638	1	1465	290	776237	non-coding	non-coding	95

Comparing Salmonella Strains

- Map Salmonella enterica Choleraesuis onto Salmonella enterica Paratypi ATCC 9150
  - Conserved genes, differing/new genes, translocations

443	-1	280986	405	outer membrane lipoprotein	88%	443	-1	289140	405	outer membrane lipoprotein
444	1	281391	118	non-coding	80%	444	1	289545	118	non-coding
445	-1	281509	816	DL-methionine transporter substrate-binding subunit	71%	445	-1	289663	816	DL-methionine transporter substrate-binding subunit
446	1	282325	38	non-coding	37%	446	1	290479	38	non-coding
447	-1	282363	654	DL-methionine transporter permease subunit	75%	447	-1	290517	639	DL-methionine transporter permease subunit
448	-1	283009	1032	DL-methionine transporter ATP-binding subunit	78%	449	-1	291163	1032	DL-methionine transporter ATP-binding subunit
449	1	284041	189	non-coding	87%	450	1	292195	189	non-coding
450	1	284230	582	D.D-heptose 1.7-bisphosphate phosphatase	66%	451	1	292384	567	D.D-heptose 1.7-bisphosphate phosphatase
451	1	284812	6165	non-coding	1%	766	1	513926	487	non-coding
451	1	284812	6165	non-coding	10%	4270	1	260436	530	non-coding
451	1	284812	6165	non-coding	14%	4290	-1	2617699	93	hypothetical protein
452	1	290977	804	2.5-diketo-D-gluconate reductase B	60%	4288	-1	2613165	804	2.5-diketo-D-gluconate reductase B
454	-1	291802	915	LysR family transcriptional regulator	67%	4286	1	2612219	915	transcriptional regulator
455	1	292717	104	non-coding	76%	4285	1	261211	104	non-coding
456	1	292821	1176	putative drug efflux protein (perhaps for chloramphenicol)	50%	4284	-1	2610959	1176	putative drug efflux protein
457	1	293997	131	non-coding	70%	4283	1	2610787	152	non-coding

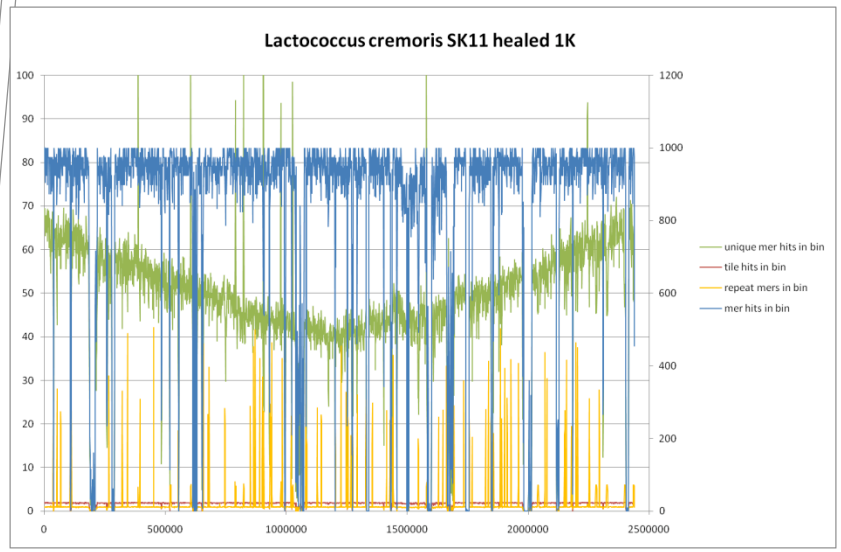
Applications Using Real Sequence Data

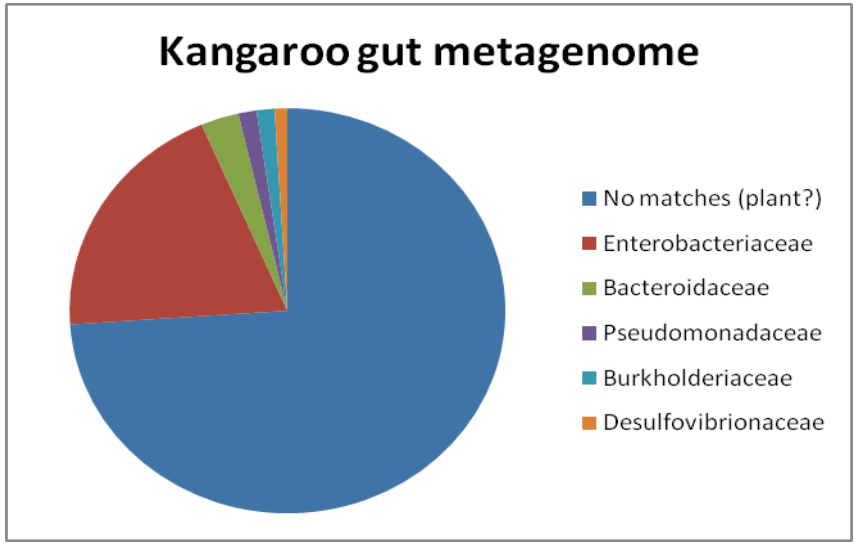
- ‘Next-gen’ sequencing machines
  - Generate gigabytes of sequence data ‘quickly’ & ‘cheaply’
  - Short strings (35 – 400bp for today)
  - Random errors at <1% rate
- Tile & map reads against all-bacteria database
  - How similar is a pure culture of an organism to references?
  - What is the composition of a mixed population from an environmental sample (metagenomics)?
  - Organism presence and gene-level sharedness counts
  - Finding regions of difference (SNPs etc)
  - Filter reads based on taxonomy
  - Runs in ‘sparse matching time’
    - 3 hours (compared to several days for conventional tools)

Sample Results

partition compared to Escherichia_coli_E24377A
Reads 80,746,662/115,646,360 25-mers from partition

seq	geneN o	geneID	geneDesc	Length	Strand	Start	mers	tiles
7	1	non-coding		335	1	1	335	1057
7	2	157158747	bifunctional aspartokinase I/homeserine dehydrogenase I	2463	1	336	1894	6022
7	3	non-coding		1	1	2799	1	5
7	4	157158096	homoserine kinase	933	1	2800	933	3909
7	5	157157451	threonine synthase	1287	1	3733	1238	7114
7	6	non-coding		289	1	5020	264	1847
7	7	157156558	hypothetical protein	498	-1	5309	413	5148
7	8	157156183	IS621. transposase	981	1	5789	981	4
7	9	non-coding		191	1	6770	168	5535
7	10	157155229	hypothetical protein	777	-1	6961	529	5785
7	11	non-coding		69	1	7738	69	233
7	12	157154855	amino acid carrier protein	1431	-1	7807	1099	7264





Problems and Extensions

- Elegance of matching code…
  - Large number of sparse lookups over large space
    - And orders of magnitude more (random) searches useful
  - Largely solved by flash etc some time soon?
  - Cleverer way of getting DB to do the matches faster?
  - Even-cleverer non-DB way of doing the matches faster?
- Scaling onto clusters and clouds
  - Matching is ‘easily’ partitionable & distributable problem
  - Provide as public service (expected by community)
- Adding more biology/capabilities
  - Amino acid codings, gene functional groups
  - Simple ‘close matching’ (SNPs, indels)
    - Additional or very different index structures
  - Finding SNPs & other differences

Humans and Similar Animals

- Thousands of human genomes coming very soon
  - 6Gig each, some internal redundancy
    - 25-mer uniqueness property still useful
  - Almost identical, but differences are the important thing
    - Point mutations, structural differences, epigenetic differences?
    - Many differences are irrelevant, but some are not
    - Very incomplete understanding of human genome
- $1000 genome coming soonish
  - Justified by medical diagnosis/prediction applications?
- What questions, what analysis, what answers?
  - Finding correlations to disease, functionality, …
  - Storing, querying and analysing deltas?
  - Need new statistical analysis techniques as well
  - Many questions but the data is coming soon

Summary

- K-mer databases useful for some purposes
  - Shared 25-mers turns out to be a useful metric
  - Sharing matrix matches taxonomic structure quite well
  - Shared 25-mers reflect shared genes
  - Applications in microbiology, metagenomics, …
- SQL queries can be used to answer some questions
  - Extensive pre-processing to populate tables first
  - Queries over all bacteria come back in ‘seconds’
- Computational challenges for some applications
  - Comparisons can be fast, but IO-limited in sparse cases
  - Low latency storage coming just in time?
- Next challenge is storing/querying human genomes
  - First flood of data is coming soon, real deluge following