TaxaSorter: A Solution to Metagenomics Projects

Li Liu

University of Florida
Interdisciplinary Center for Biotechnology Research (ICBR)

June 20th, 2007

454 Metagenomics @ UF

454 Projects @ UF

- Acquired 454 Sequencer two years ago
- Only have a handful of regular genome projects
- Majorities are
  - transcriptome sequencing, followed by custom microarray design
  - amplicon sequencing
  - metagenomics projects

Metagenomics Projects @ UF

- biodiversity studies
- Bacteria genome sequencing with high contaminations
- Pathogen identification in environmental samples

Metagenomics Questions

Questions metagenomics projects try to answer

- Biodiversity related questions:
  - how many species reside in the sampled environment
  - their relative abundances
- Comparisons related questions:
  - differences / similarities among sample environments
  - the importance / significance of those differences

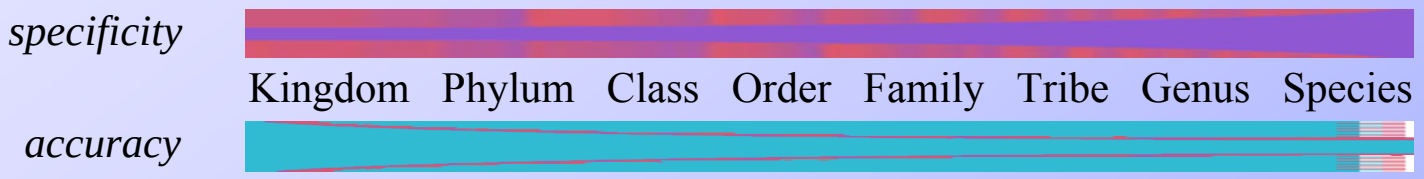
The fundamental question is

- assign the correct taxonomic origin to each and every sequence.

UNIVERSITY OF FLORIDA ICBR

What's the taxonomic origin of a sequence?

BUT, ON WHICH RESOLUTION LEVEL?



- Resolution:
  - at which taxonomy level are we seeking answers
  - the accuracy rates highly depend on the taxonomy levels

TaxaSorter Analysis Methods

Two Basic Approaches

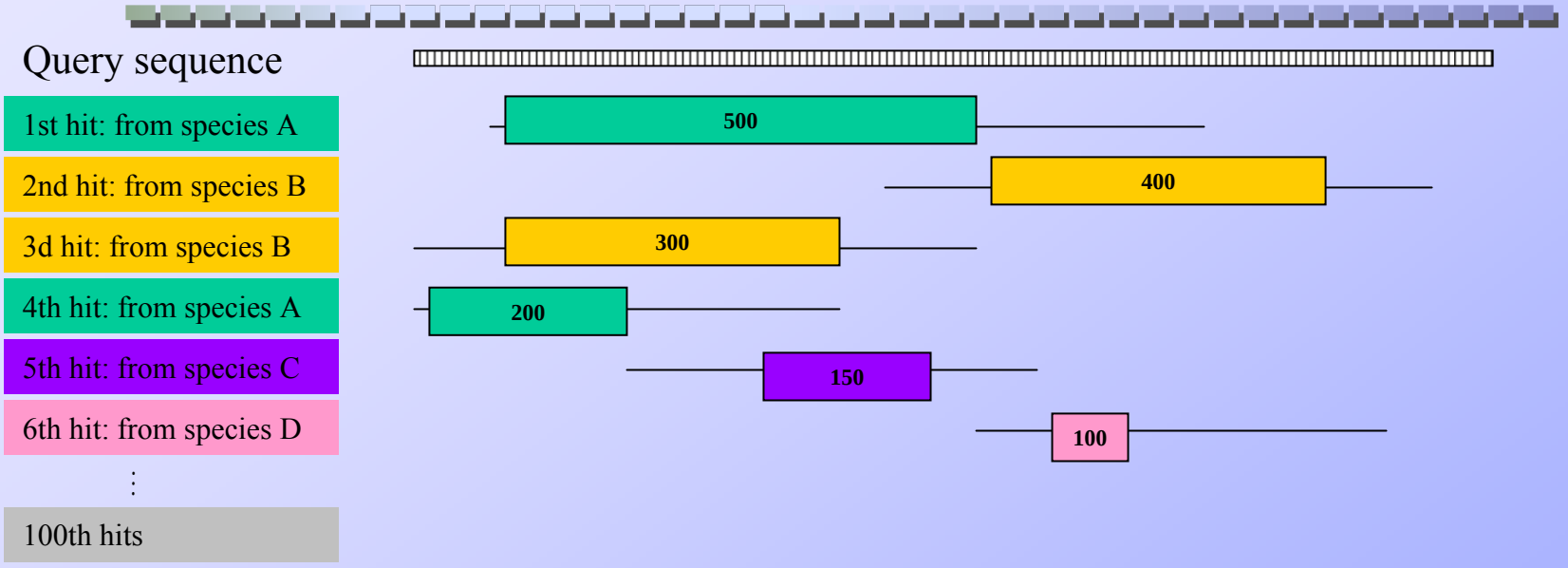
- Similarity searches
  - BLAST, etc.
- Genomic signature
  - N-nucleotide motifs & Pattern Recognition Algorithm

BLAST assisted decision making

Accuracy of BLAST assisted decision making depends on the following factors

- database content:
  - coverage: full coverage vs. partial coverage vs. zero coverage
  - selected conserved genes (e.g. 16s rRNA) vs. all genes (NR/NT)
  - redundancy: high vs. medium vs. non-redundance
- top N hits: is the first hit always the best hit?
- cutoff values
  - e-value, bit-score, identity, etc.
- programs
  - blastn vs. blastx vs. tblastx

TaxaSorter Algorithm



Step 1

```
for (each identified species) {
        find non-redundant aligned regions;
        calculate a normalized score (n-score) based on these regions;
}
```

Take care of the problem of database redundancy and segmented sequences from the same genome

Step 2

Calculate the probability that a species is the real origin {

$$p_i = \frac{\text{n-score}_i}{\sum \text{n-score}}$$

}

Under the assumption that this query sequence must come from one of the species represented in top 100 hits.

Step 3

Sort into the NCBI taxonomy trees.

Help address the specificity vs. accuracy problem.

UNIVERSITY OF FLORIDA ICBR

Comparing two taxonomy trees

Step 1:
Traverse tree A through all nodes (taxa).
For (each taxon) {
    record the number of sequences linked or sub-linked to this node;
}

Step 2:
Traverse tree B through all nodes (taxa).
For (each taxon) {
    record the number of sequences linked or sub-linked to this node;
}

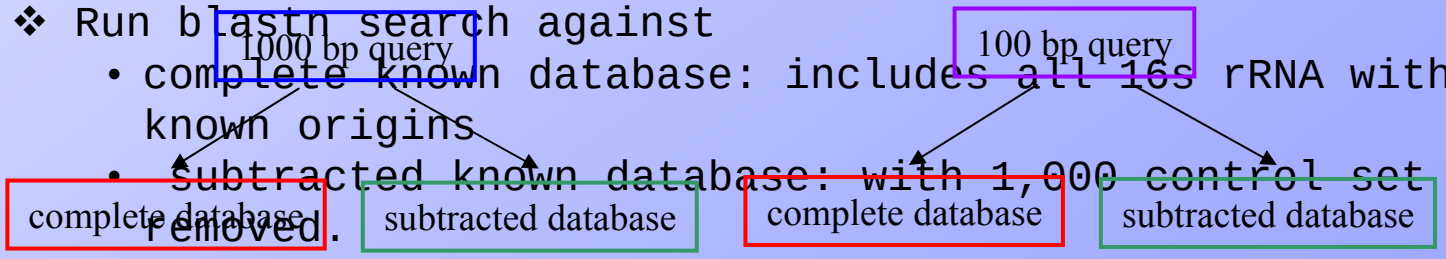
Step 3:
For (each taxon) {
    run "Fisher Exact Test" to test if this taxon is significantly over-represented or under-represented in tree A compared to tree B.
}

Step 4:
Correct for multiple test errors (Family-wise step-down False Discovery Rate)

Simulations

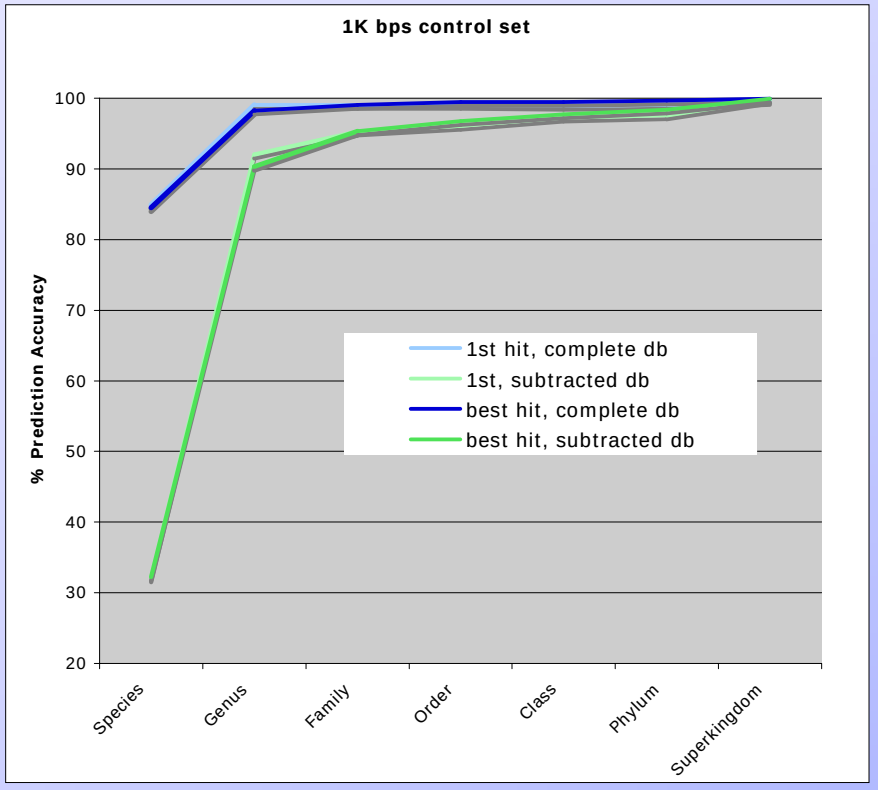
RDP (Ribosomal Database Project) v9.45: 16s rRNA

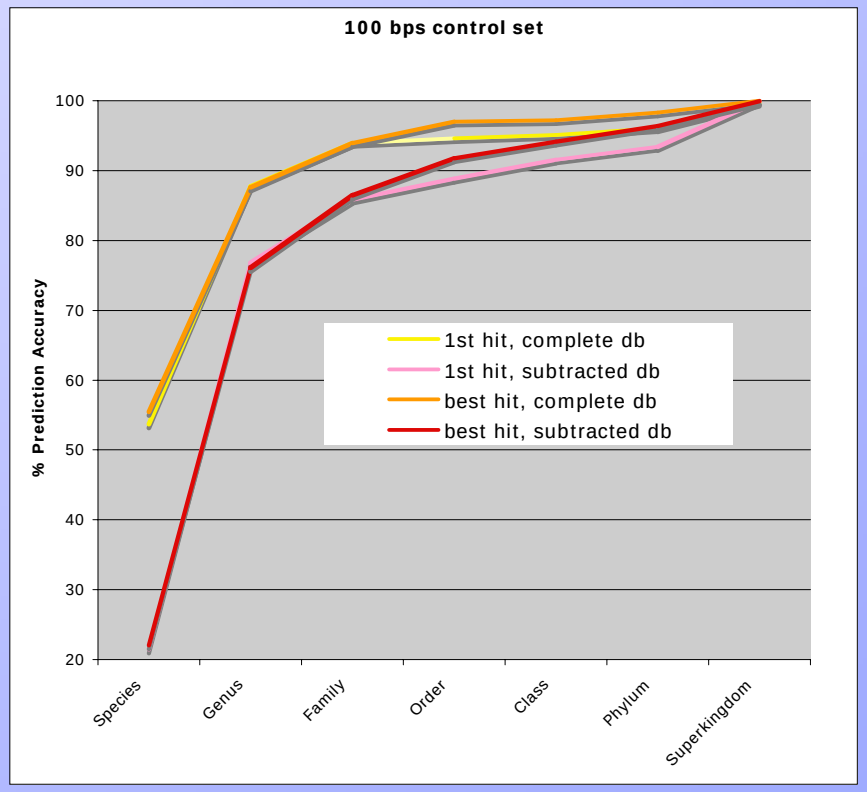
- Use only 16s rRNA genes with known origin
  - by removing all “uncultured” or “unidentified” sequences
- Randomly select 1,000 16s rRNAs with known origins as control set.
  - the average sequence length of the control set is 1,000 bps, which represents the simulation of Sanger sequences.
  - randomly select a segment of 100 bps from each control sequence to simulate 454 GS20 sequences.
- Run blastn search against
  - complete known database: includes all 16s rRNA with known origins
  - subtracted known database: with 1,000 control set removed.



Simulations

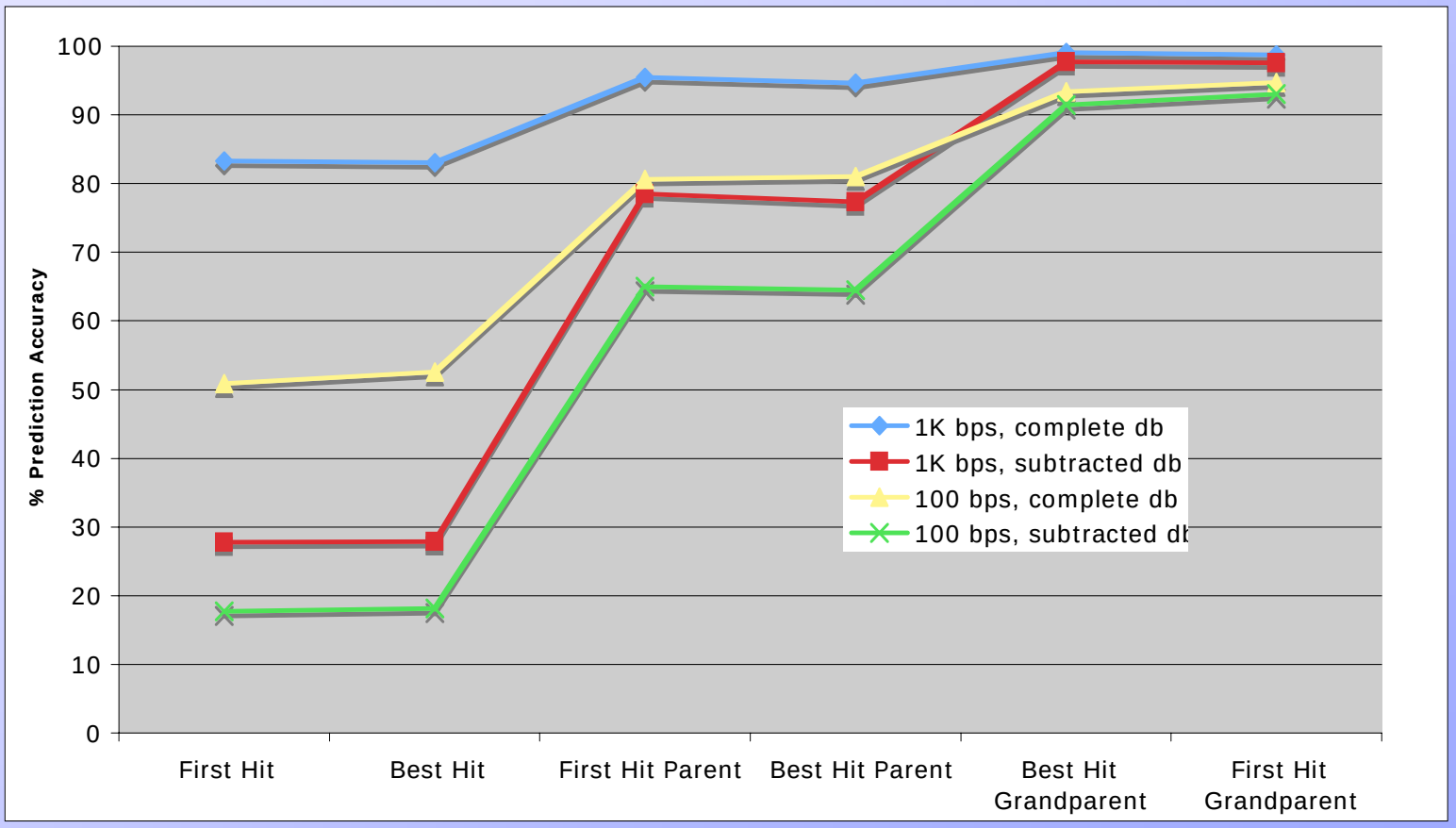
Specificity & Accuracy





Simulations

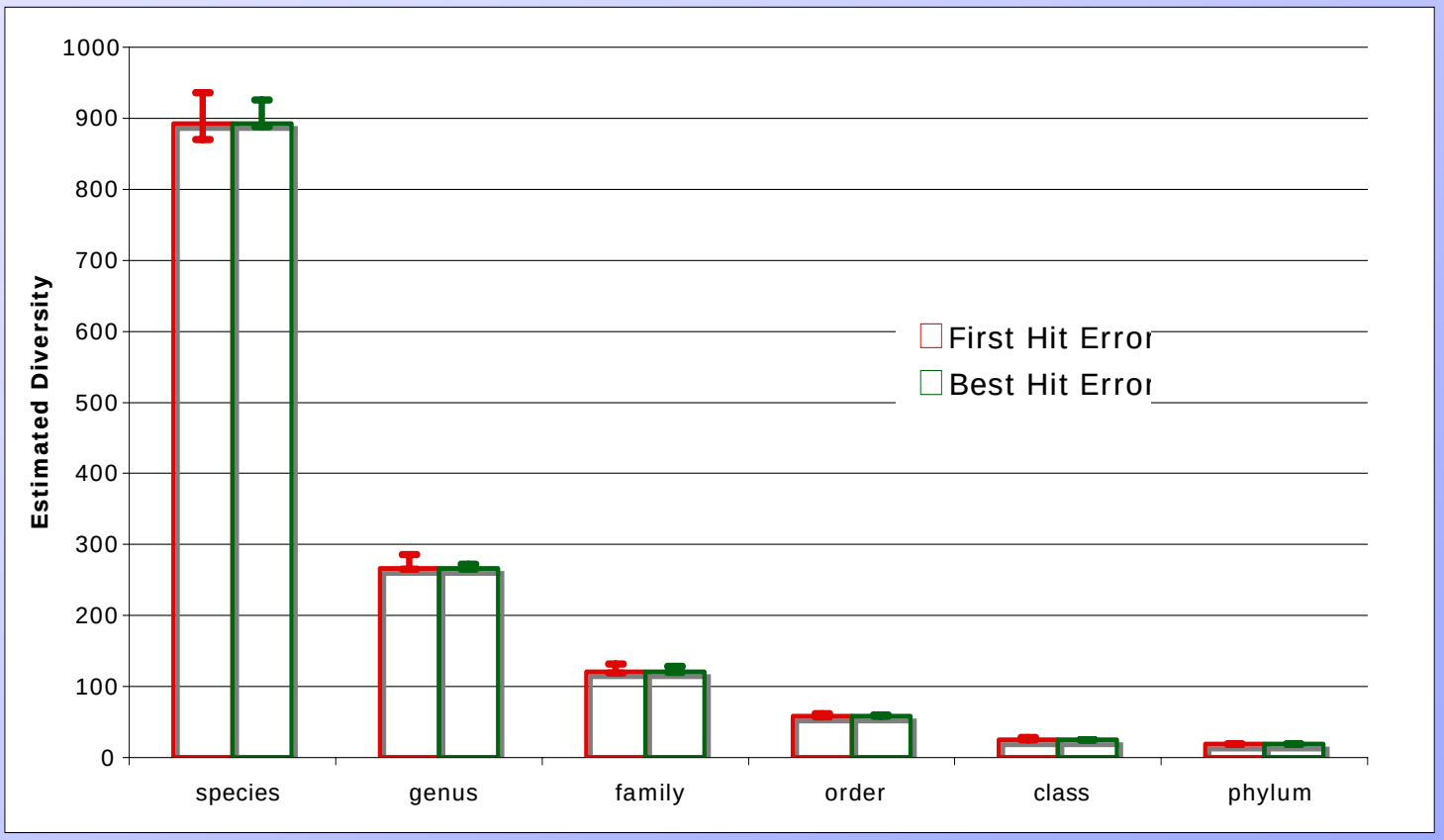
Accuracy on the lowest taxonomy levels and two levels



UNIVERSITY OF FLORIDA ICBR

Simulations

Biodiversity Estimation



Simulations Conclusion

Complete Database vs. Subtracted Database

- On the extremes, using complete database improves prediction accuracy with up to 50%

1K bps vs. 100 bps

- Longer sequences help improve prediction accuracy with up to 30%

Best Hit vs. First Hit

- If the sequence is long, 95% of first hits are best hits.
- If the sequence is short, only 70% of first hits are best hits.

Higher Taxonomy Level vs. Lower Taxonomy Level

- With longer sequence searched against complete database, prediction on species level is trustworthy.
- With shorter sequences searched against incomplete database, prediction on genus or family level is trustworthy.

Tree Comparisons: compared to real origins

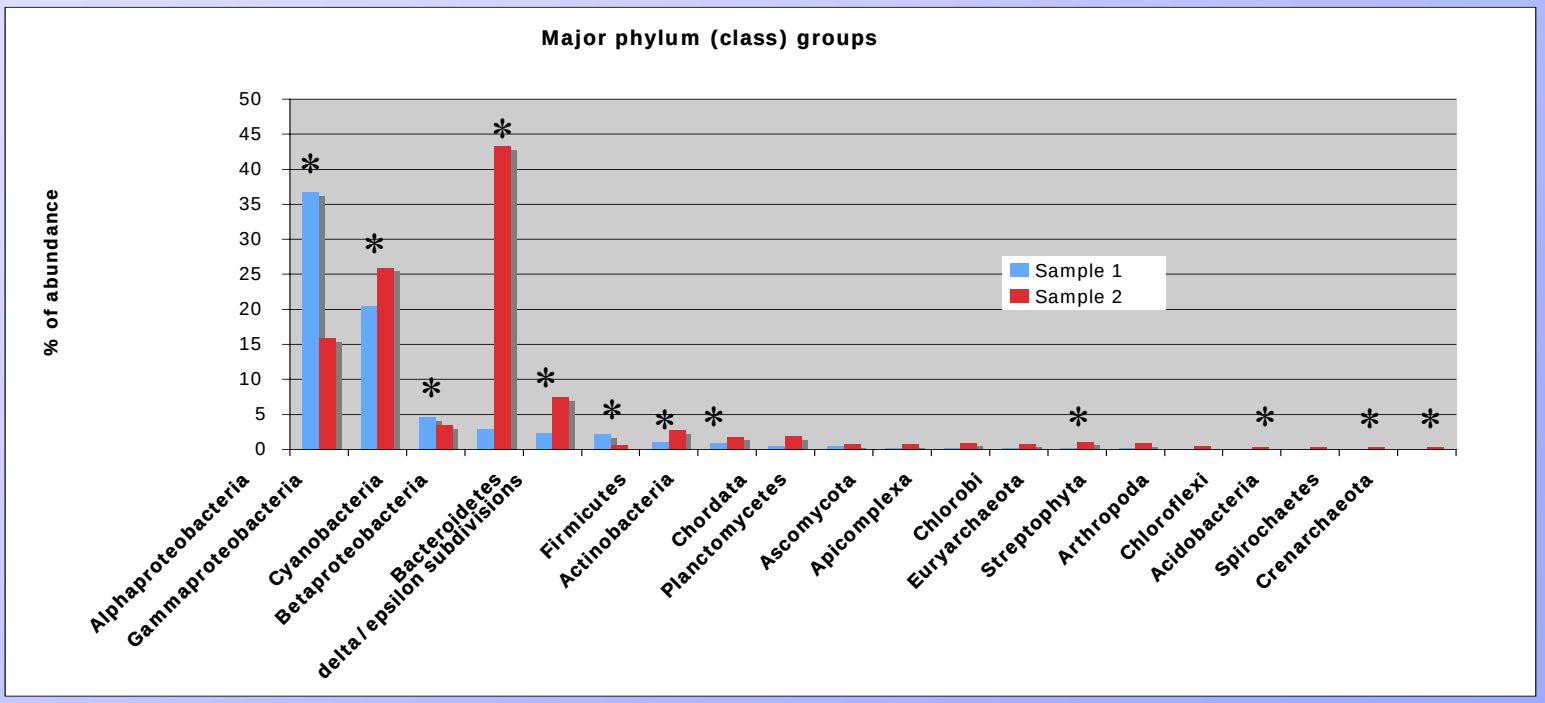
- No significant difference was found for any taxon above "family" level for any of the simulation conditions at $p <= 0.05$ level.

UNIVERSITY OF FLORIDA ICBR

Sargasso Sea Metagenomic Project

Biodiversity (10K Sanger reads for

	species	genus	family	order	class	phylum
Sample 1	440 - 556	266 - 352	154 - 237	99 - 158	48 - 68	25 - 32
Sample 2	527 - 732	342 - 506	195 - 345	124 - 212	63 - 87	29 - 37



UNIVERSITY OF FLORIDA ICBR

TaxaSorter as a module in BlastQuest system

root , cellular organisms , Bacteria

- Proteobacteria (8594, 8669, 7355.44)

  + Gammaproteobacteria (2595, 2630, 3171.4)
  + Betaproteobacteria (4328, 4349, 1881.65)
  + Alphaproteobacteria (1594, 1621, 1920.31)
  + unclassified Proteobacteria (10, 7, 29.01)
  + delta/epsilon subdivisions (60, 55, 345.93)
  + environmental samples (7, 7, 7.15)

UNIVERSITY OF FLORIDA ICBR

TaxaSorter as a module in BlastQuest system

Blast Information (Best Hit) for Tax 'Alphaproteobacteria' (TaxID=28211) (Project ID: 20031664)

Query Def	Query_len	All_Hits	Tax ID	Hit Probablity	Hit_num	Hit_accession	Hsp_evalue	Hsp_bit_score	Hsp_identity	Hsp_align	Hit_def
SLAPE30TR	848	Details	335992	0.05	1	AAZ21542	0.0	323.168	162	193	ribonucleotide reductase alpha subunit [Can ubique HTCC1062] >gil71083426|ref|YP_ reductase alpha subunit [Candidatus Pelagib HTCC1062]
SSABZ81TF	868	Details	335992	0.02	1	AAZ21982	0.0	341.273	163	210	phosphate ABC transporter [Candidatus Pel HTCC1062] >gil71083866|ref|YP_266586 transporter [Candidatus Pelagibacter ubique
SLBDG36TF	849	Details	335992	0.06	1	AAZ20858	0.0	257.299	122	154	Cell division protein MraZ [Candidatus Pela HTCC1062] >gil71082742|ref|YP_265461 SAR11_0033 [Candidatus Pelagibacter ubi >gil91207205|sp|Q4FPN3|MRAZ_PELUB >gil91762836|ref|ZP_01264801.1| hypothet PU1002_06186 [Candidatus Pelagibacter u >gil91718638|gb|EAS85288.1| hypothetical [Candidatus Pelagibacter ubique HTCC100
SHAAE43TF	870	Details	335992	0.03	1	AAZ21032	0.0	420.239	219	290	lysine-tRNA ligase [Candidatus Pelagibacte >gil71082916|ref|YP_265635.1| lysyl-tRNA Pelagibacter ubique HTCC1062]
SKAOJ31TR	900	Details	335992	0.03	1	AAZ21521	0.0	304.294	164	262	replicative DNA helicase (dnaB) [Candidat HTCC1062] >gil71083405|ref|YP_266124 helicase (dnaB) [Candidatus Pelagibacter ub >gil91762172|ref|ZP_01264137.1| replicativ [Candidatus Pelagibacter ubique HTCC100 >gil91717974|gb|EAS84624.1| replicative D [Candidatus Pelagibacter ubique HTCC100
SLADG57TF	925	Details	335992	0.03	1	AAZ21340	0.0	230.335	111	145	protein release factor Class I (RF-1) [Candi HTCC1062] >gil71083224|ref|YP_265943 Class I (RF-1) [Candidatus Pelagibacter ubi
SLBAC62TR	805	Details	314232	0.03	1	ZP_01004140	0.0	223.402	105	161	putative transport system permease protein [SKA53] >gil84509250|gb|EAQ05709.1| pu permease protein [Loktanella vestfoldensis

TaxaSorter in BlastQuest - 1st Hit Selected

Blast Information for Query 'SLAPE30TR' (QueryLen: 848 Project ID: 20031664)

Hit_num	Hsp_num	Hit_accession	Hsp_evalue	Hsp_bit_score	Hsp_identity	Hsp_align	Query_from	Query_to	Hit_from	Hit_to	Hit_def
1	1	AAZ21542	0.0	323.168	162	193	33	611	81	273	ribonucleotide reductase alpha subunit [Candidatus Pelagibacter ubique HTCC1062] >gi|71083426|ref|YP_266145.1| ribonucleotide reductase alpha subunit [Candidatus Pelagibacter ubique HTCC1062]
2	1	ZP_01264116	0.0	323.168	162	193	33	611	81	273	ribonucleotide reductase alpha subunit [Candidatus Pelagibacter ubique HTCC1002] >gi|91717953|gb|EAS84603.1| ribonucleotide reductase alpha subunit [Candidatus Pelagibacter ubique HTCC1002]
3	1	ABG32698	0.0	193.741	105	194	27	578	940	1133	ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent [Roseobacter denitrificans OCh 114]
4	1	ZP_01017542	0.0	191.43	98	166	81	578	944	1105	ribonucleotide-diphosphate reductase alpha subunit [Parvularcula bermudensis HTCC2503] >gi|84690148|gb|EAQ15989.1| ribonucleotide-diphosphate reductase alpha subunit [Parvularcula bermudensis HTCC2503]
5	1	ZP_00962544	0.0	189.889	104	192	27	572	940	1131	ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent [Sulfitobacter sp. NAS-14.1] >gi|83841768|gb|EAP80937.1| ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent [Sulfitobacter sp. NAS-14.1]
6	1	ZP_00955043	0.0	189.889	104	192	27	572	940	1131	ribonucleotide-diphosphate reductase alpha subunit [Sulfitobacter sp. EE-36] >gi|83846675|gb|EAP84551.1| ribonucleotide-diphosphate reductase alpha subunit [Sulfitobacter sp. EE-36]
7	1	ZP_01001471	0.0	188.734	102	195	27	578	941	1135	ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent [Oceanicola batsensis HTCC2597] >gi|84388198|gb|EAQ01150.1| ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent [Oceanicola batsensis HTCC2597]

UNIVERSITY OF FLORIDA ICBR

Our Group

Bill Farmerie

Fahong Yu

Yijun Sun

Jason Li

Aaron Gardner

Li Liu
liliu@biotech.ufl.edu

UNIVERSITY OF FLORIDA ICBR