Novel algorithms for assembling and searching protein sequences in metagenomic datasets

Shibu Yooseph

Professor

Informatics Department

J. Craig Venter™ INSTITUTE

Metagenomics

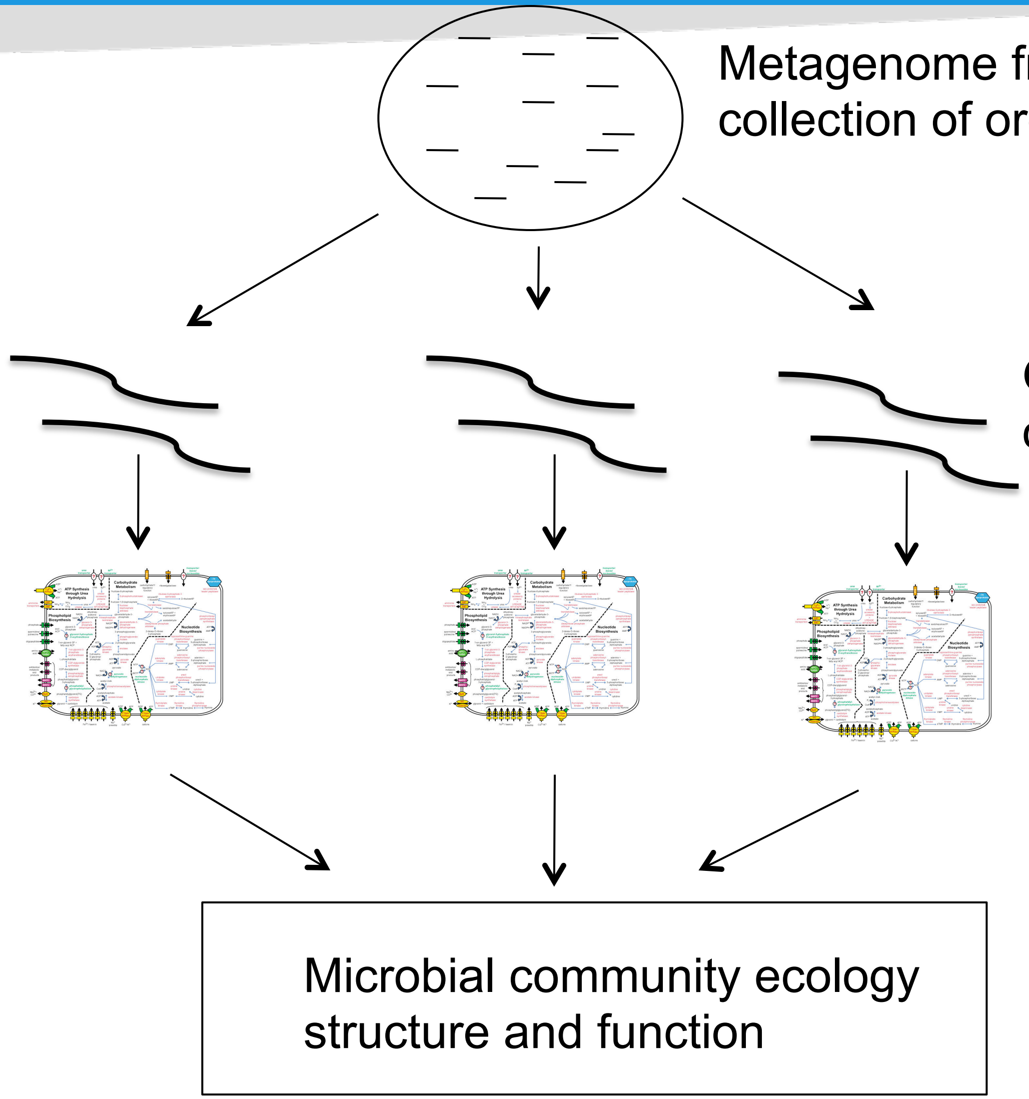
- Examining genomic content of organisms in community/environment to better understand
  - Diversity of organisms
  - Their roles and interactions in the ecosystem

- Cultivation independent approach to study microbial communities
  - DNA directly isolated from environmental sample and sequenced

J. Craig Venter™
INSTITUTE

Metagenomics

Sequence Assembly and Searching



Metagenome from collection of organisms

Genomes of constituent organisms

Metabolic and functional reconstruction

Microbial community ecology structure and function

Microbial communities

- Collection of organisms (taxonomically distinct)
- Varying abundances
- (Possibly) different %GC content and codon usage biases
- Strain variants, genome rearrangements, etc.

- Community complexity is a function of carbon and energy sources, and environmental variables like temperature, pH, salinity, etc.

Metagenomic assembly

- Inference of complete (or near complete) genomes of constituent microbial species from the sequenced DNA sample

- Metagenomic assembly of even medium complexity microbial communities is a challenge (*Rusch et al 2007*, *Qin et al 2010*)
  - Fragmented assemblies with short contigs
  - Large proportion of the input nucleotide reads remain unassembled

Metagenomic assembly

- Assemblies serve as substrate for annotation of genome features and downstream functional analysis

- Consequence of poor assemblies
  - Fragmentary gene sequences
  - Annotation of fragmentary sequences can suffer from lack of accuracy and specificity
  - Annotation and analysis restricted to assembled data leads to an incomplete picture of the microbial community

Read-based analysis

- Next Generation Sequencing (NGS) technologies produce vast amounts of sequence data
  - For instance, one run of Illumina HiSeq 2500 can generate 6 billion paired-end reads (100 bp)

- Annotation of all reads computationally prohibitive
  - Also suffers from lack of accuracy due to short read lengths

Functional analysis revisited

Our goal: *(Yang and Yooseph, NAR 2013)*

Reconstruction of (near) full-length protein sequences from their constituent peptide fragments identified on short read data

Inference of complete protein sequences from metagenomic data sets should provide a more accurate picture of the functional and metabolic potential of the microbial community

Why this approach could work

1. High coding density (~90%) in prokaryotic and viral genomes

   - Majority of nucleotide reads will contain portion of a gene

Why this approach could work

2. There are de novo gene finders for metagenomic data that can predict genes on short reads with high accuracy and are computationally efficient; for instance MetaGeneAnnotator (MGA) (*Noguchi et al 2008*) and FragGeneScan (FGS) (*Rho et al 2010*)

   ○ Can predict fragmentary protein sequences (short peptides) from nucleotide reads rapidly

Why this approach could work

3. Amino acid conservation extends over a larger taxonomic range compared with nucleotide conservation

   ○ Thus, nucleotide polymorphisms, a striking feature of natural microbial populations and a major confounding factor in nucleotide assembly of related strains, will not be an obstacle when the assembly is carried out at the amino acid level, as there is a high degree of protein sequence conservation across strains from the same species

Peptide Assembly Framework (*Yang and Yooseph 2013*)

1. **Gene Finding (GF)**
   - Use a gene finder to identify (partial) protein-coding genes (*short peptides*) from reads

2. **Assembly (SPA)**
   - Construct a de Bruijn graph *G* on the set of peptides obtained in GF stage
   - Traverse *G* in an informed fashion (using *k-mer* coverage and read overlap) to identify probable paths that correspond to proteins

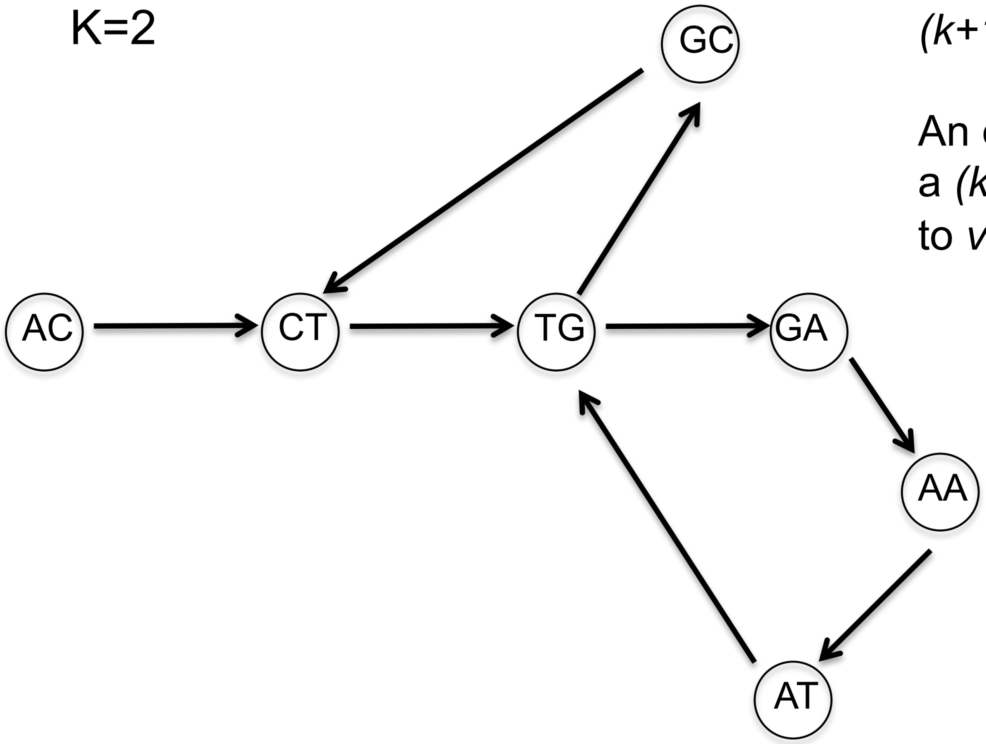
3. **Post processing (PP)**
   - Refine sequence set identified in SPA using corresponding gene finder, to generate final set of assembled sequences

de Bruijn graphs or k-mer graphs

Constructed on sequences

ACTGAATGCT

K=2



A de Bruijn graph *G* is a directed graph:

The vertices in *G* denote the distinct *k-mers* (that is, substrings of length *k*) present in the sequences

The (directed) edges in *G* represent the distinct *(k+1)-mers* present in the sequences

An edge exists from vertex v_i to vertex v_j if *S* has a *(k+1)-mer* whose length *k* prefix corresponds to v_i and whose length *k* suffix corresponds to v_j.

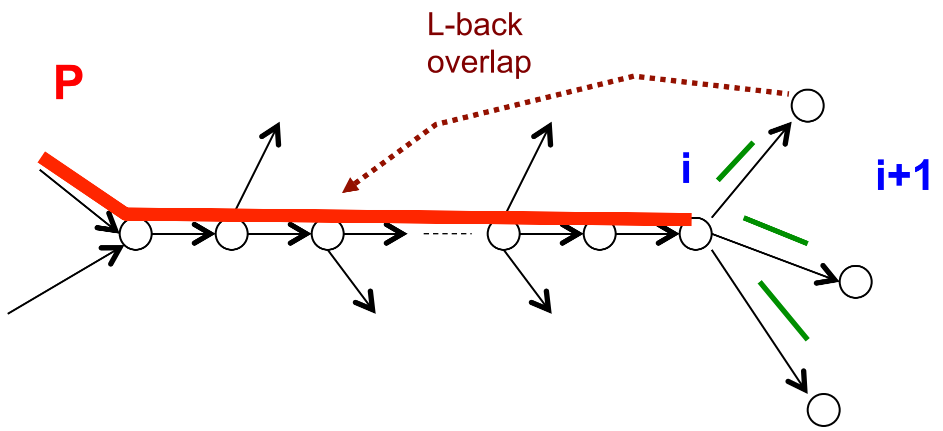
J. Craig Venter™ INSTITUTE

de Bruijn graph

- Allows for compact representation of read overlap information
- Used in many nucleotide assemblers (*Idury and Waterman, 1995; Pevzner et al 2001*)
- Provides alternative framework compared to overlap-layout-consensus approach
  - Primary approach for most NGS data

- *For our peptide assembly framework, we construct de Bruijn graphs using the amino acid alphabet*

Graph traversal:
Identification of Initial Path Set (IPS)



S: set of all reads (short peptide sequences)
G_i: set of reads that can thus-far be fully placed on P
B_i: set of reads that only partially overlap with P

Node i+1 is one with thickest extension
G_{i+1} and B_{i+1} derived from G_i and B_i

Path extension termination:
Node *i* is terminal node, L-back overlap below threshold
Repeat handling fails

Subsequent steps

- Merge highly similar paths
- Recruitment of unassigned sequences
- Extension and merging of paths

- Post-processing step to handle over-prediction by gene-finders

Implementation

- ***SPA output***:
  - Sequences
  - MSA of its constituent peptide fragments
  - Various statistics on the path, including path length, depth of coverage at each alignment column, and the entropy of each column, are also output
- *Implementation:*
  - C++
- Availability:
  - http://sourceforge.net/projects/spa-assembler

J. Craig Venter™ INSTITUTE

Evaluation

- Performance compared against alternate strategy of *assembling nucleotide reads and identifying genes on the resulting contigs*

- Six different nucleotide assemblers were used in the evaluation
  - Velvet (*Zerbino and Birney, 2008*)
  - CLC (*www.clcbio.com*)
  - SOAPdenovo (*Li et al 2010*)
  - MetaVelvet (*Namiki et al, 2011*)
  - Meta-IDBA (*Peng et al 2011*)
  - IDBA-UD (*Peng et al 2012*)

J. Craig Venter™
INSTITUTE

Evaluation

- Specificity, Sensitivity, Chimera rate, and Read Assembly rate

- Let *P* denote the set of amino acid sequences output by a method and let *R* denote the set of reference protein sequences
- A sequence in *P* is defined to be *c% length matched* to a sequence in the reference set *R* if the two sequences have an alignment with ≥90% sequence identity and the alignment covers ≥*c*% of the length of the reference sequence.

Specificity (at *c%*) = $\frac{number\ of\ sequences\ in\ \mathcal{P}\ that\ are\ c\%\ length\ matched}{total\ number\ of\ sequences\ in\ \mathcal{P}}$

Sensitivity (at *c%*) = $\frac{number\ of\ sequences\ in\ \mathcal{R}\ that\ are\ c\%\ length\ matched}{total\ number\ of\ sequences\ in\ \mathcal{R}}$

Datasets

- Amino acid sequence sets

  *To evaluate SPA **algorithm** only*

  DS1. Individual genomes

  DS2. Protein fragments from a collection of genomes

- Nucleotide sequence sets

  *To evaluate SPA **framework***

  DS3. (Simulated oral microbiome)

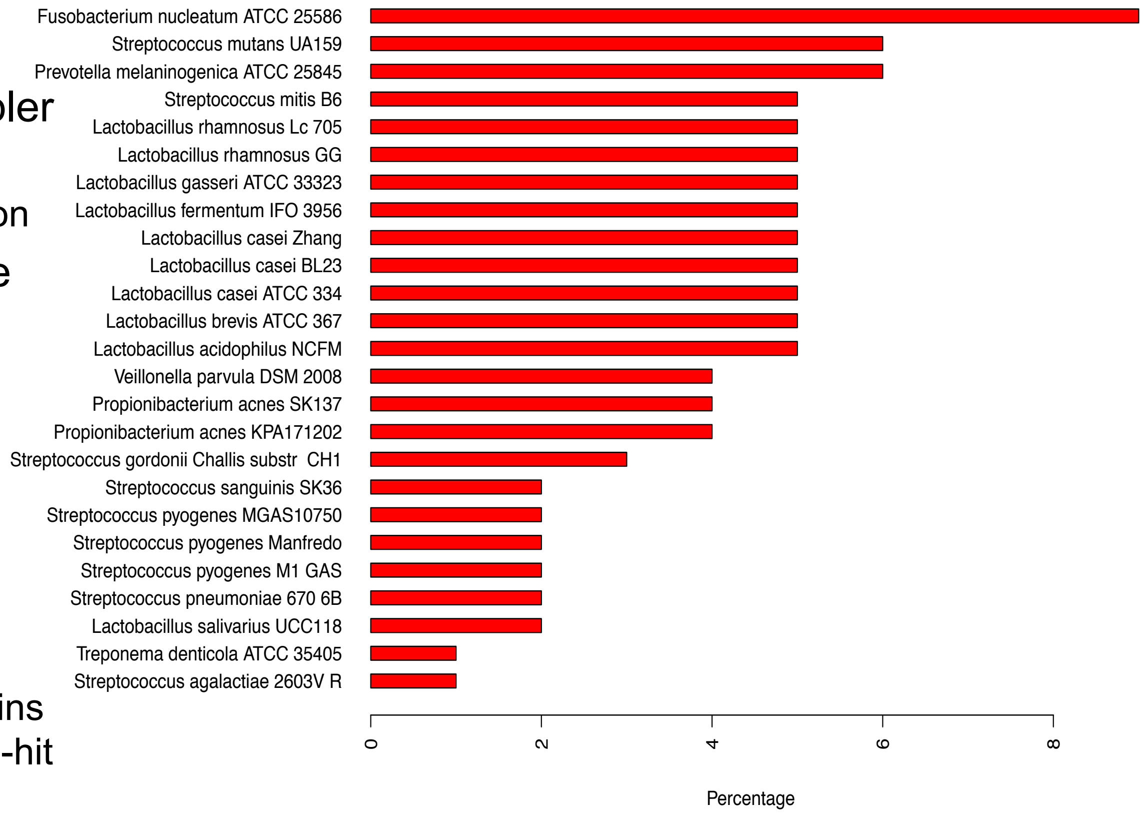
  DS4. (Simulated marine metagenome)

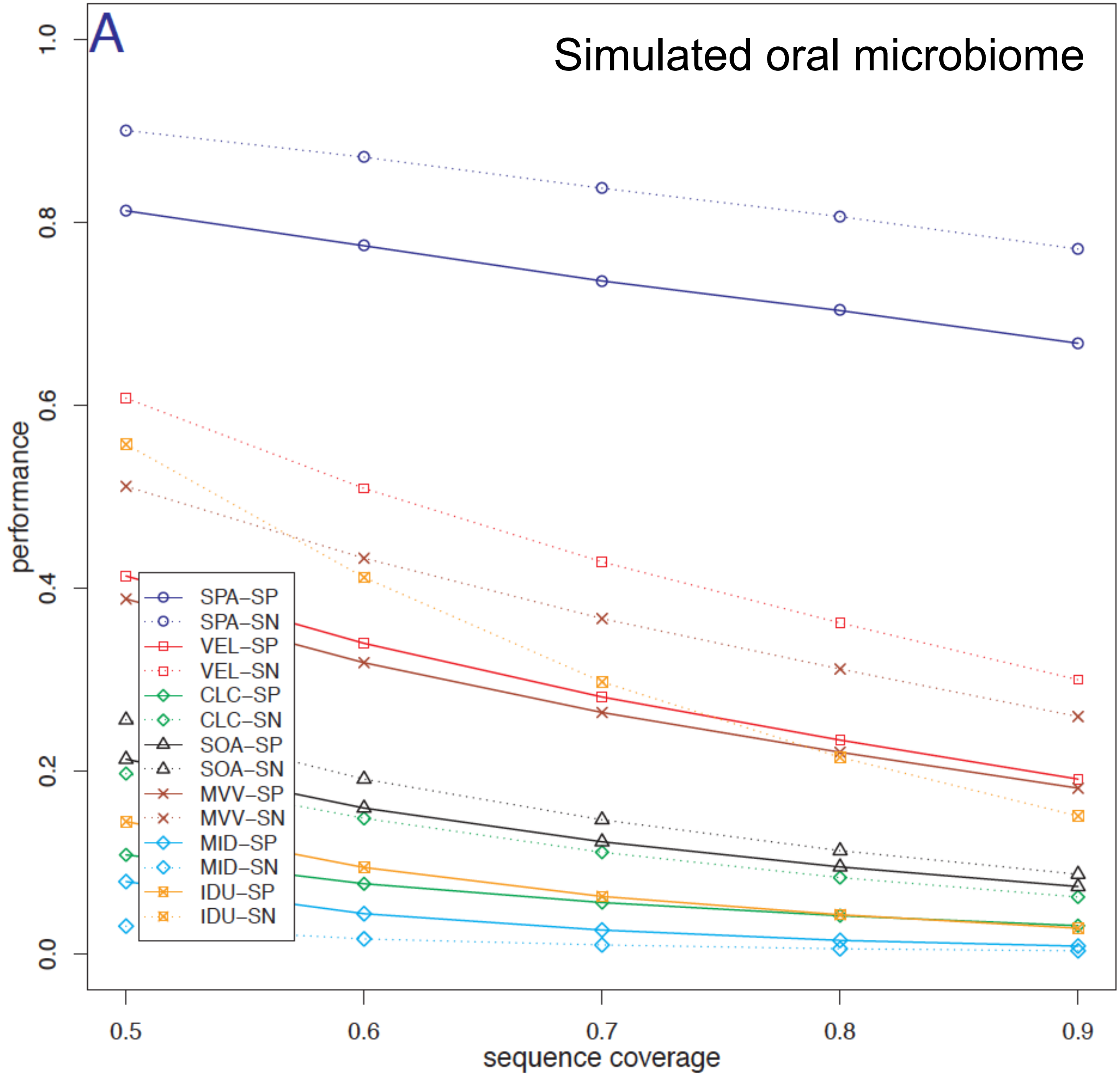
  DS5. (HMP Saliva sample)

  DS6. (HMP Stool sample)

DS3. Simulated oral microbiome

- Initial set of 25 genomes
- Generated a community of 500 genomes using the population sampler in MetaSim (*Richter et al 2008*)
  - Jukes-Cantor model of DNA evolution
- These 500 genome sequences were then sampled (at 10X depth of coverage) using wgsim
  - Generate 100 bp paired-end reads from inserts of size 300 bp
- Total 115,991,500 reads
- The reference protein set
  - 40,724 non-redundant sequences
  - clustering the combined set of proteins from the initial 25 genomes using cd-hit at 95%

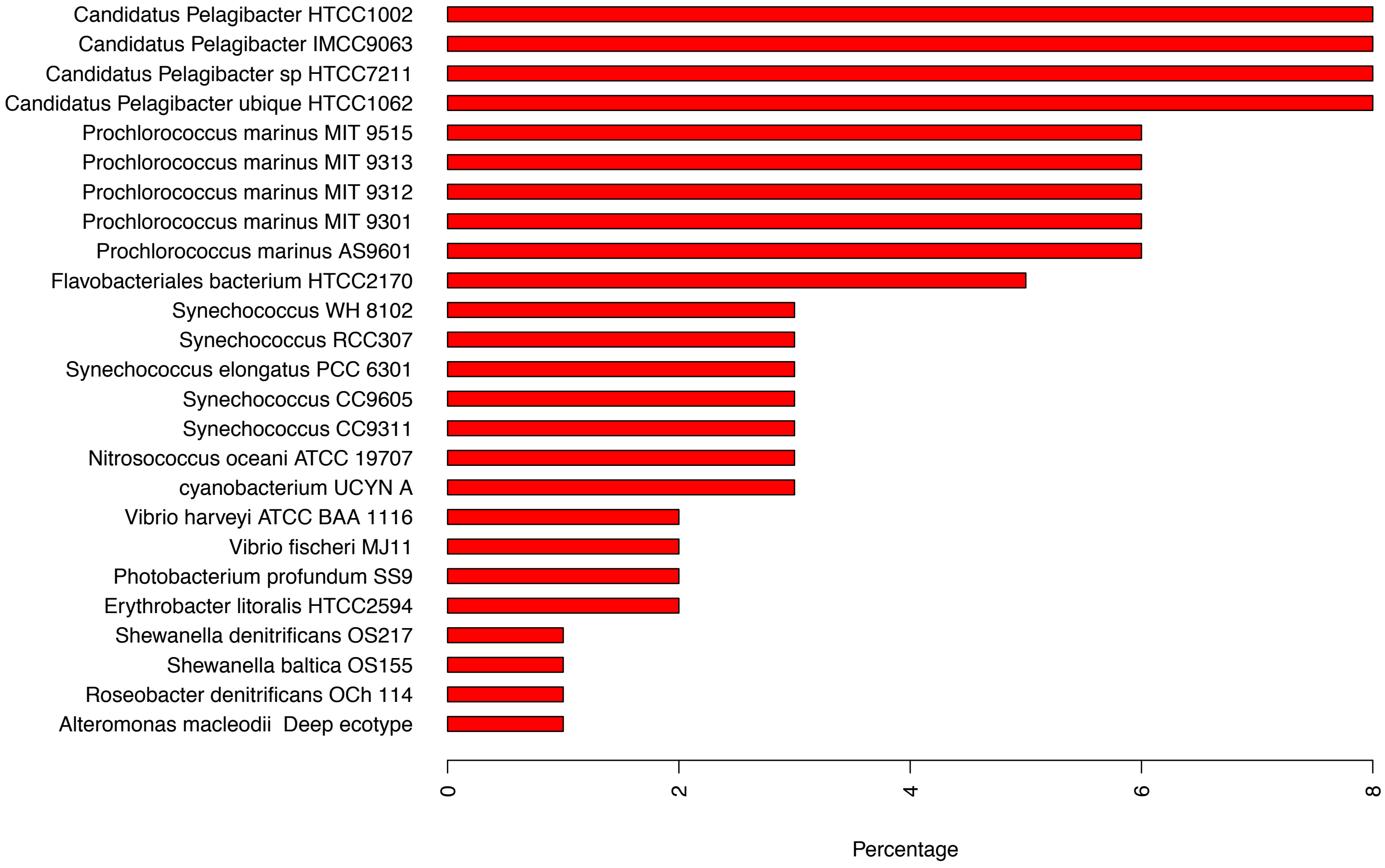




Simulated oral microbiome

0% sequence error
FGS gene-finder

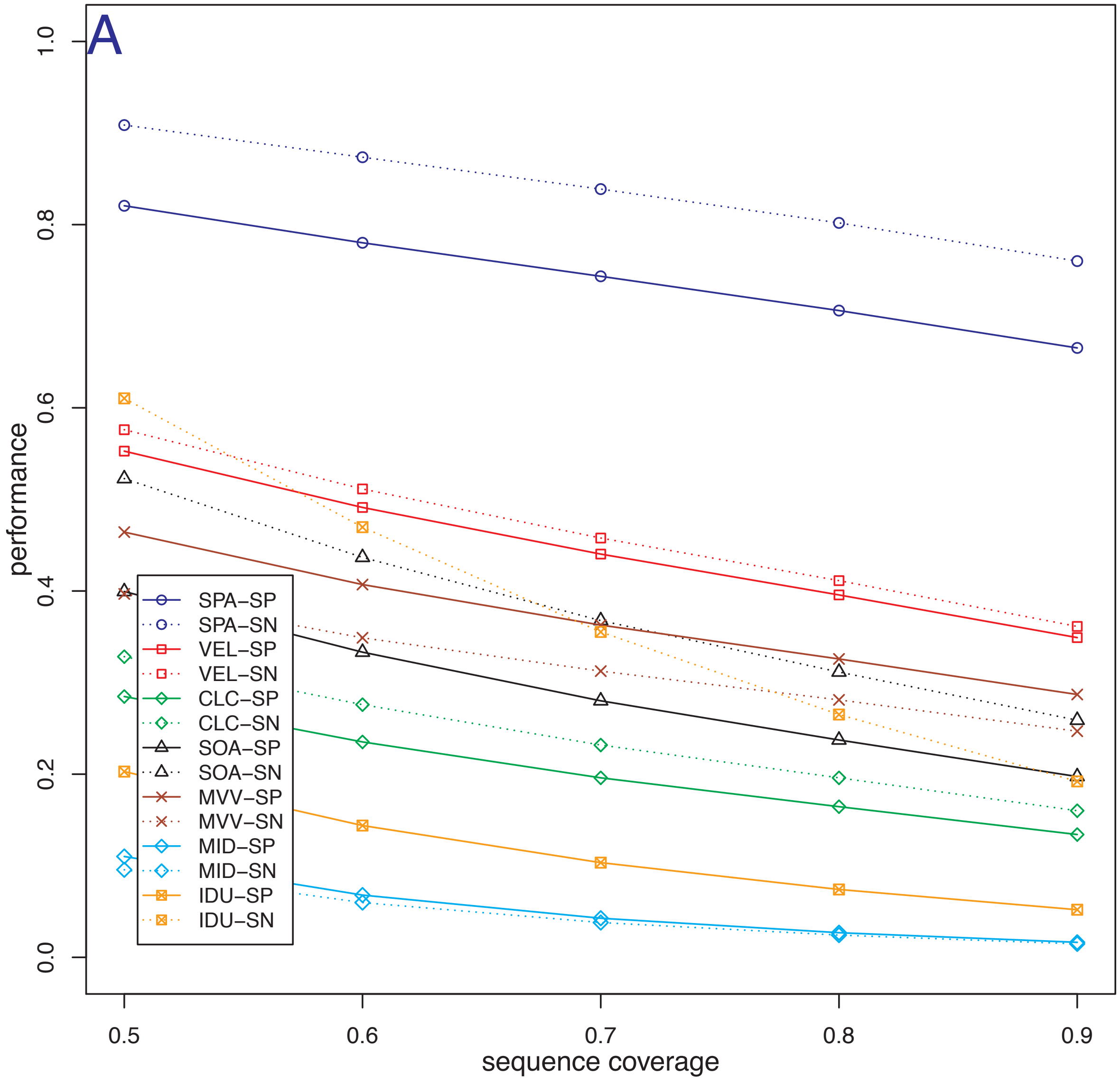
DS4. Simulated marine metagenome



- A total of 103,915,150 reads were generated in a manner similar to the method used for DS3
- The reference protein set
  - 64,913 non-redundant sequences

J. Craig Venter™ INSTITUTE

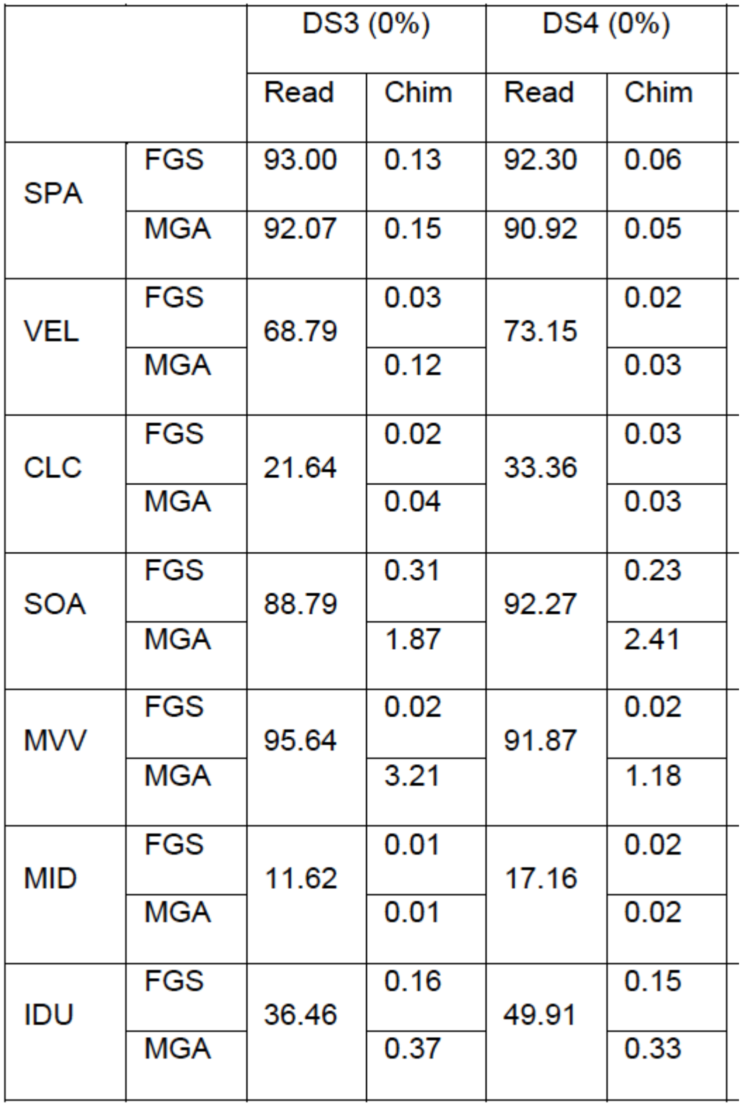
Simulated marine metagenome



0% sequence error
FGS gene-finder

J. Craig Venter™ INSTITUTE

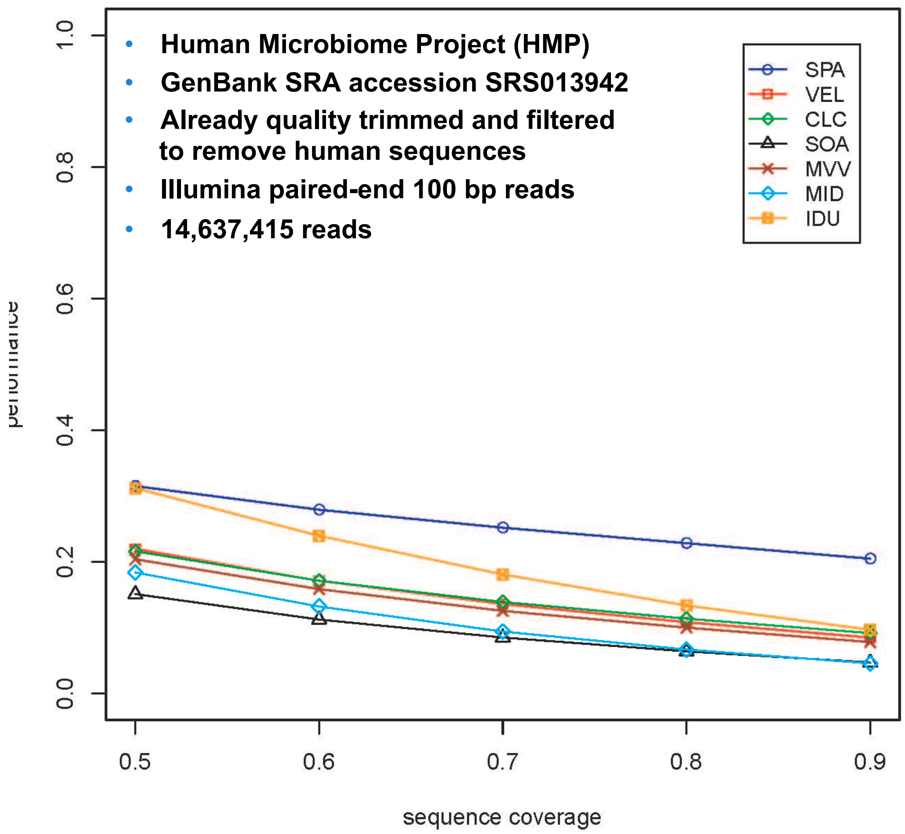
Read assembly rate and Chimera rate



		DS3 (0%)		DS4 (0%)	
		Read	Chim	Read	Chim
SPA	FGS	93.00	0.13	92.30	0.06
	MGA	92.07	0.15	90.92	0.05
VEL	FGS	68.79	0.03	73.15	0.02
	MGA		0.12		0.03
CLC	FGS	21.64	0.02	33.36	0.03
	MGA		0.04		0.03
SOA	FGS	88.79	0.31	92.27	0.23
	MGA		1.87		2.41
MVV	FGS	95.64	0.02	91.87	0.02
	MGA		3.21		1.18
MID	FGS	11.62	0.01	17.16	0.02
	MGA		0.01		0.02
IDU	FGS	36.46	0.16	49.91	0.15
	MGA		0.37		0.33

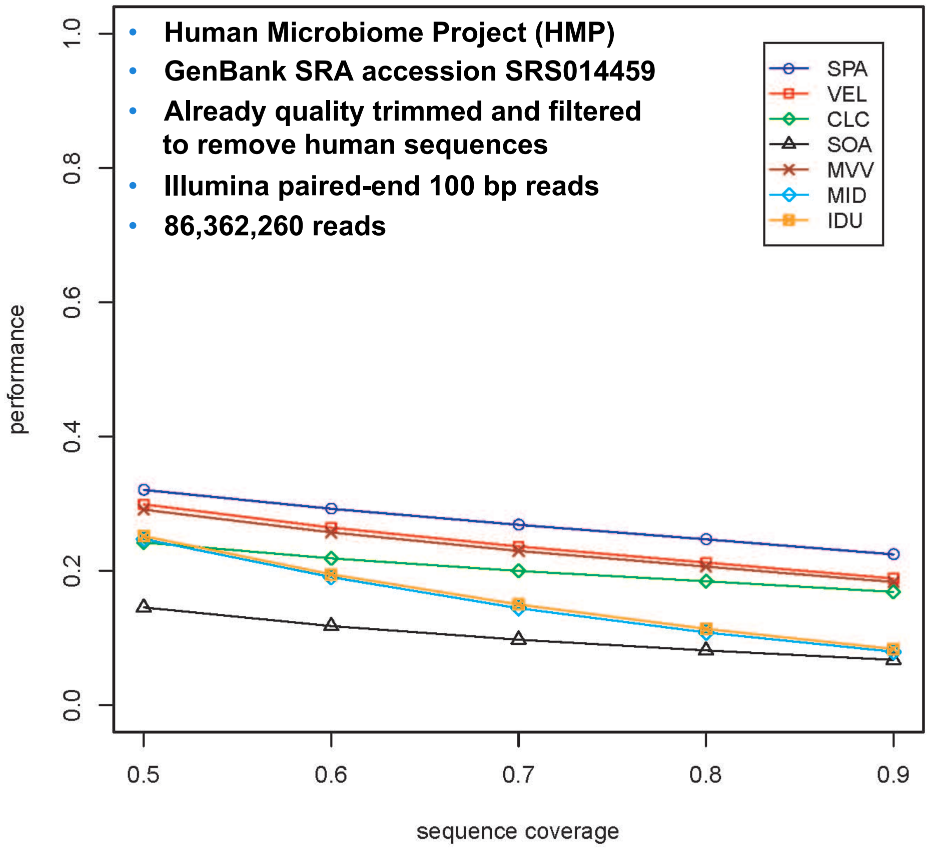
HMP data sets

Saliva



- Human Microbiome Project (HMP)
- GenBank SRA accession SRS013942
- Already quality trimmed and filtered to remove human sequences
- Illumina paired-end 100 bp reads
- 14,637,415 reads

Stool



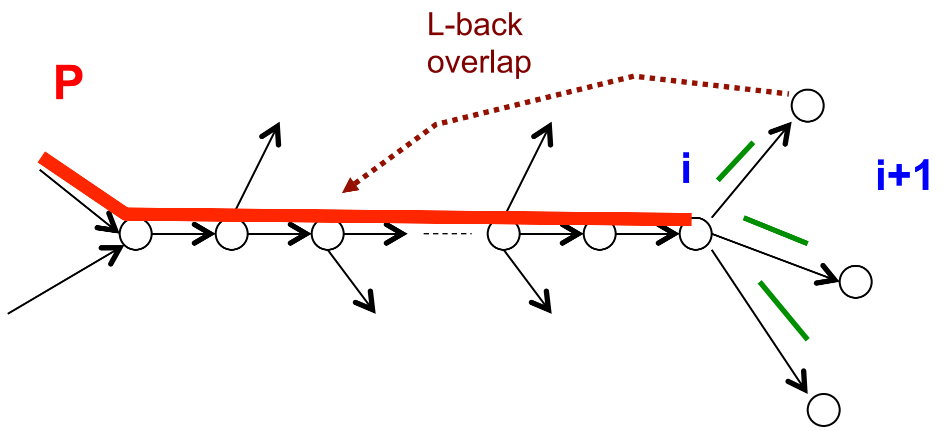
- Human Microbiome Project (HMP)
- GenBank SRA accession SRS014459
- Already quality trimmed and filtered to remove human sequences
- Illumina paired-end 100 bp reads
- 86,362,260 reads

FGS gene-finder

Evaluation summary

- For protein reconstruction, SPA framework performs much better than alternate nucleotide assembly based approach
- Low chimera rates (for all methods)
- SPA has amongst highest read assembly rate
- SPA performance using FGS slightly better than that using MGA

Graph traversal: Identification of Initial Path Set (IPS)

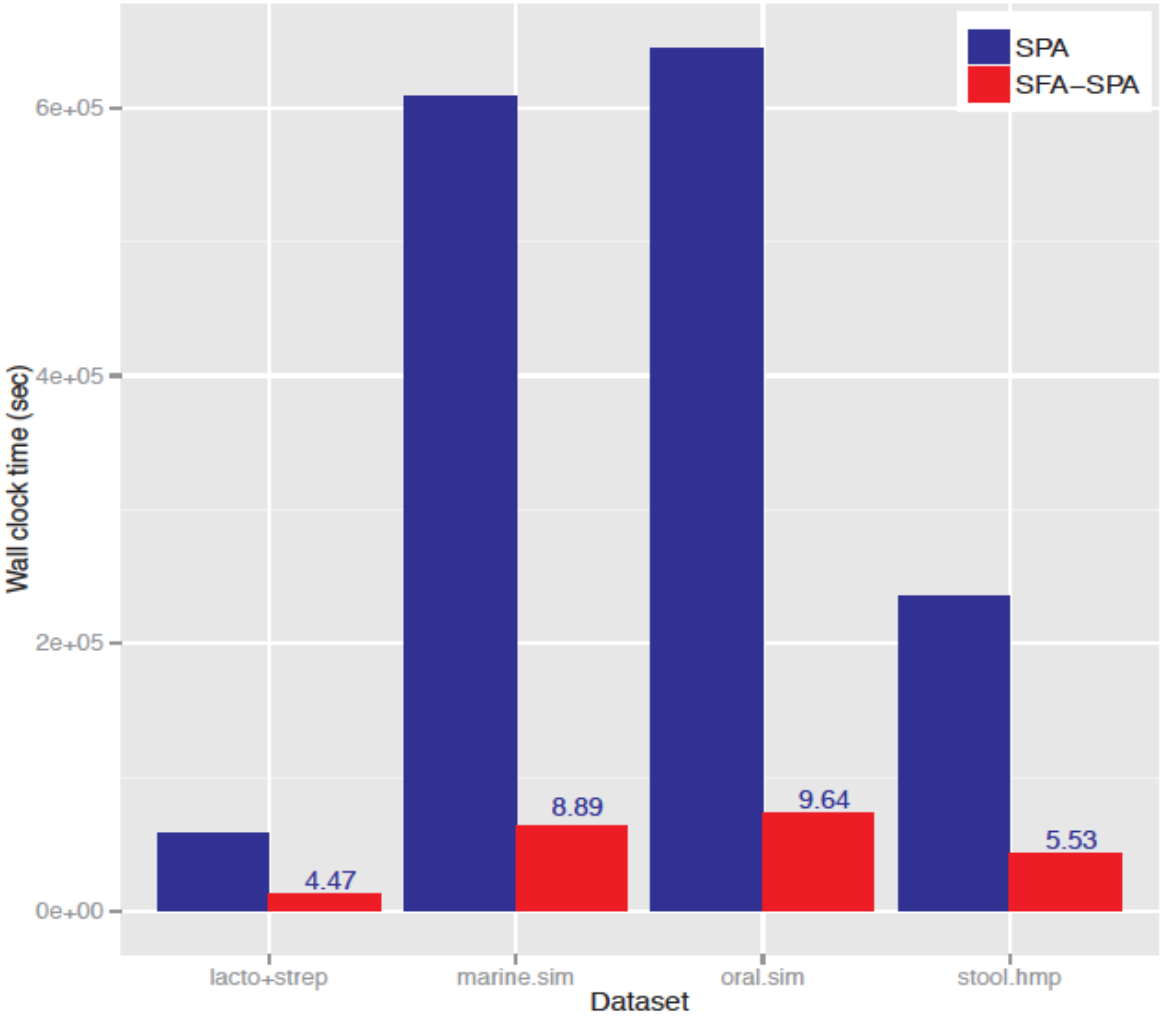


S: set of all reads (short peptide sequences)
G_i: set of reads that can thus-far be fully placed on P
B_i: set of reads that only partially overlap with P

Node i+1 is one with thickest extension
G_{i+1} and B_{i+1} derived from G_i and B_i

Speed-up using Suffix Array

(*Yang, Zhong, and Yooseph, in prep*)



J. Craig Venter™ INSTITUTE

Search problem

Input:

- Query (or reference) protein sequence *Q*
- Database *db* of protein sequences

Goal:

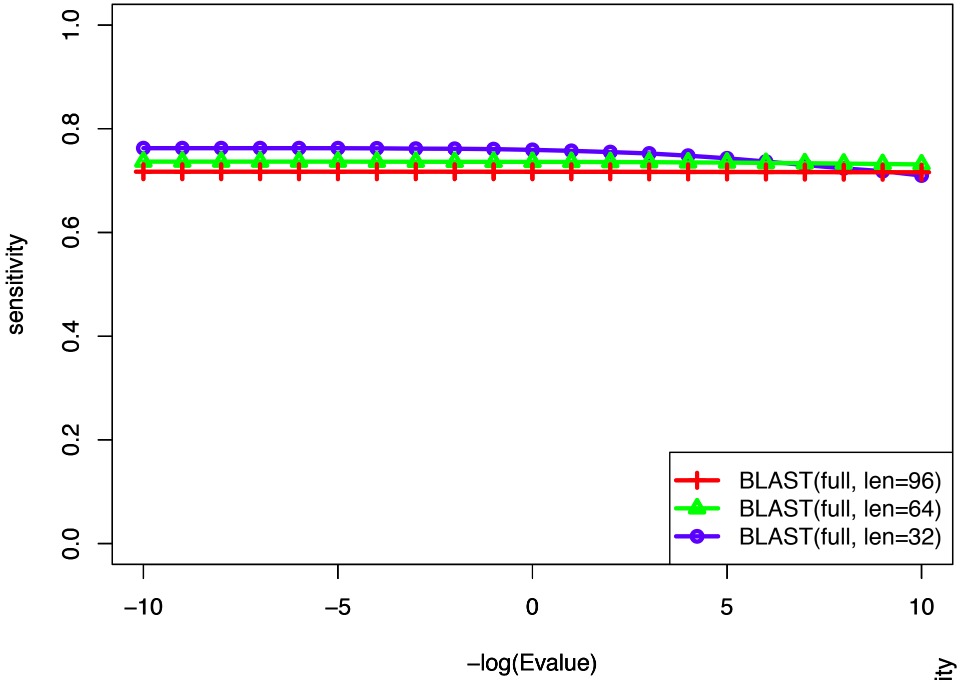
Identify sequences in *db* that are homologous to *Q*

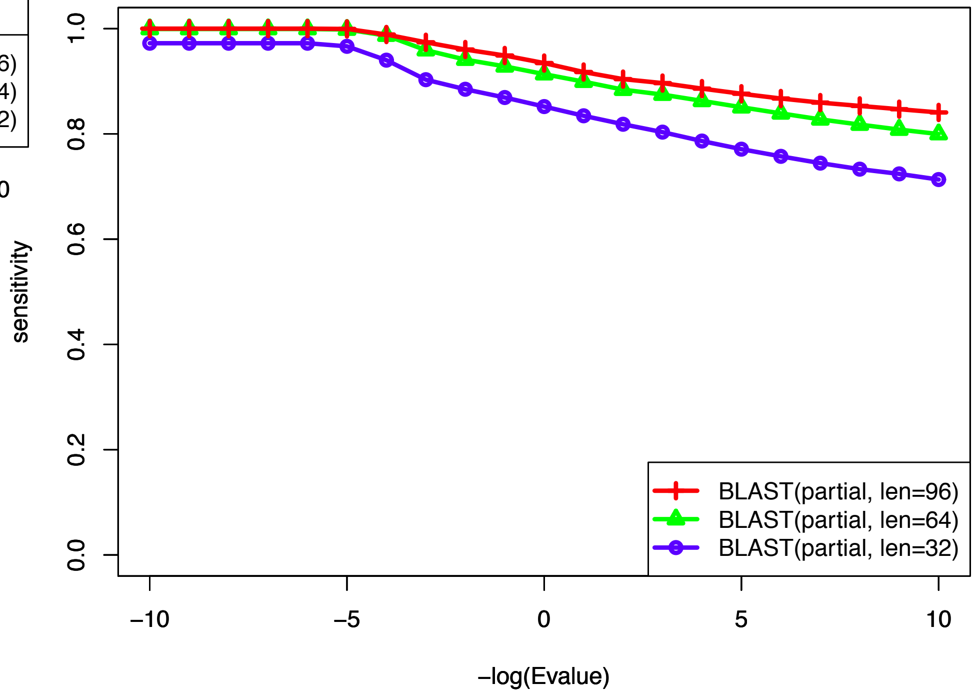
Popular solution: BLAST

- What happens when *db* contains **short** peptide sequences?
  - Collection of gene predictions (mostly fragmentary) from a metagenomic dataset

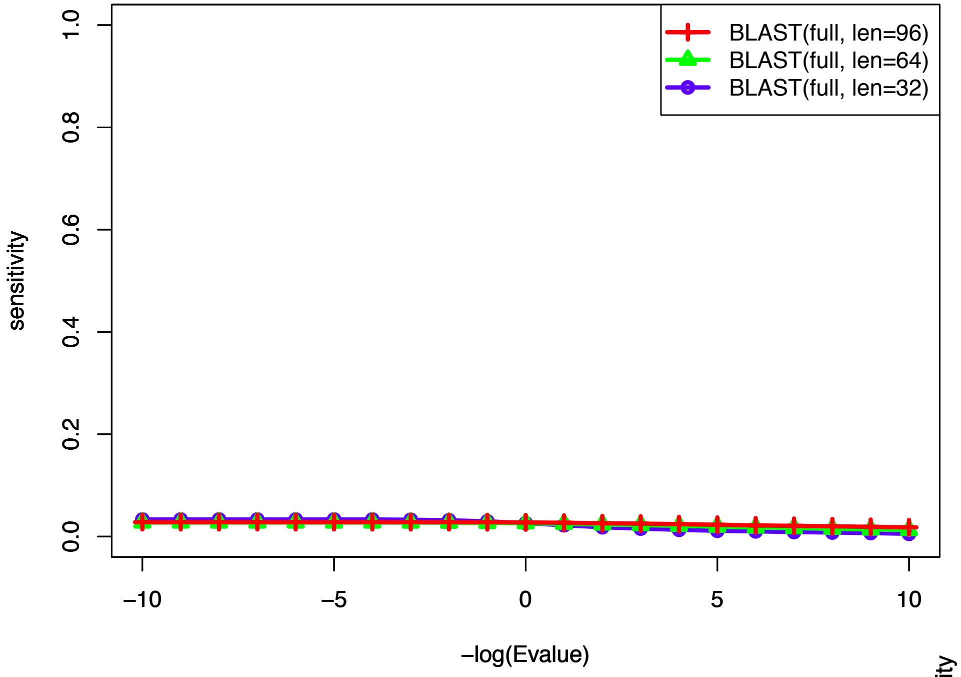
- Performance of BLAST would be dependent on
  - Length of sequences in *db*
  - Degree of conservation of protein family

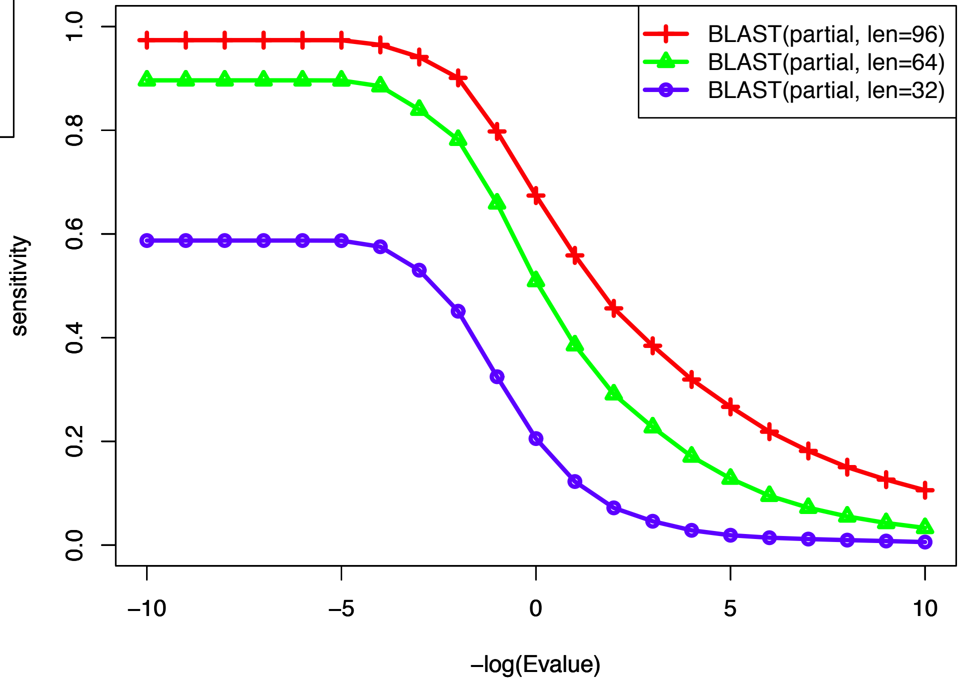
Example: RNA polymerase beta subunit (PF04563)





Example: LigT like Phosphoesterase (PF02834)





Can we do better?

What if there are sequences in *db* that are from the same (or related) protein family as *Q* ?

- As is the case for metagenomic data

Thus, what if, while searching *db*, we also assemble overlapping *db* sequences related to *Q*?

With assembled sequences, improved ability to detect HSPs, and therefore improved sensitivity of identification of homologs of Q

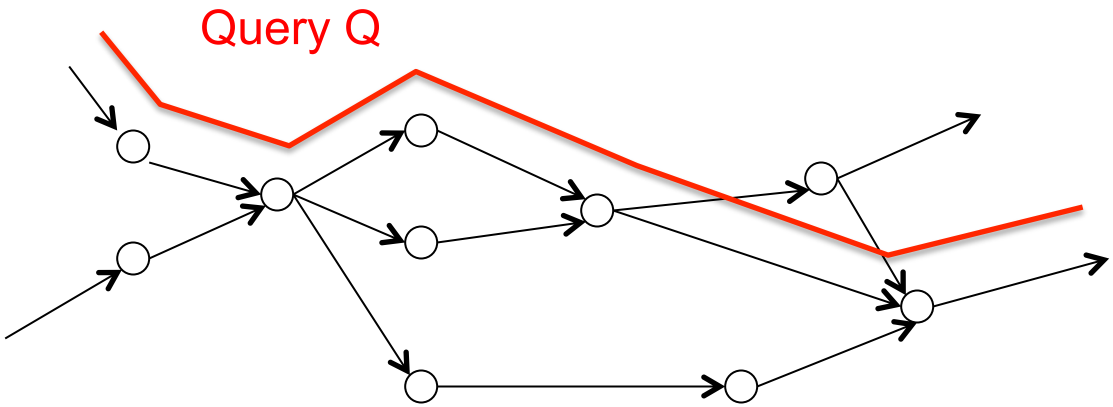
GRASP

Guided Reference based Assembly of Short Peptides

(*Zhong, Yang, and Yooseph, in prep*)

Conceptual idea using k-mer graph

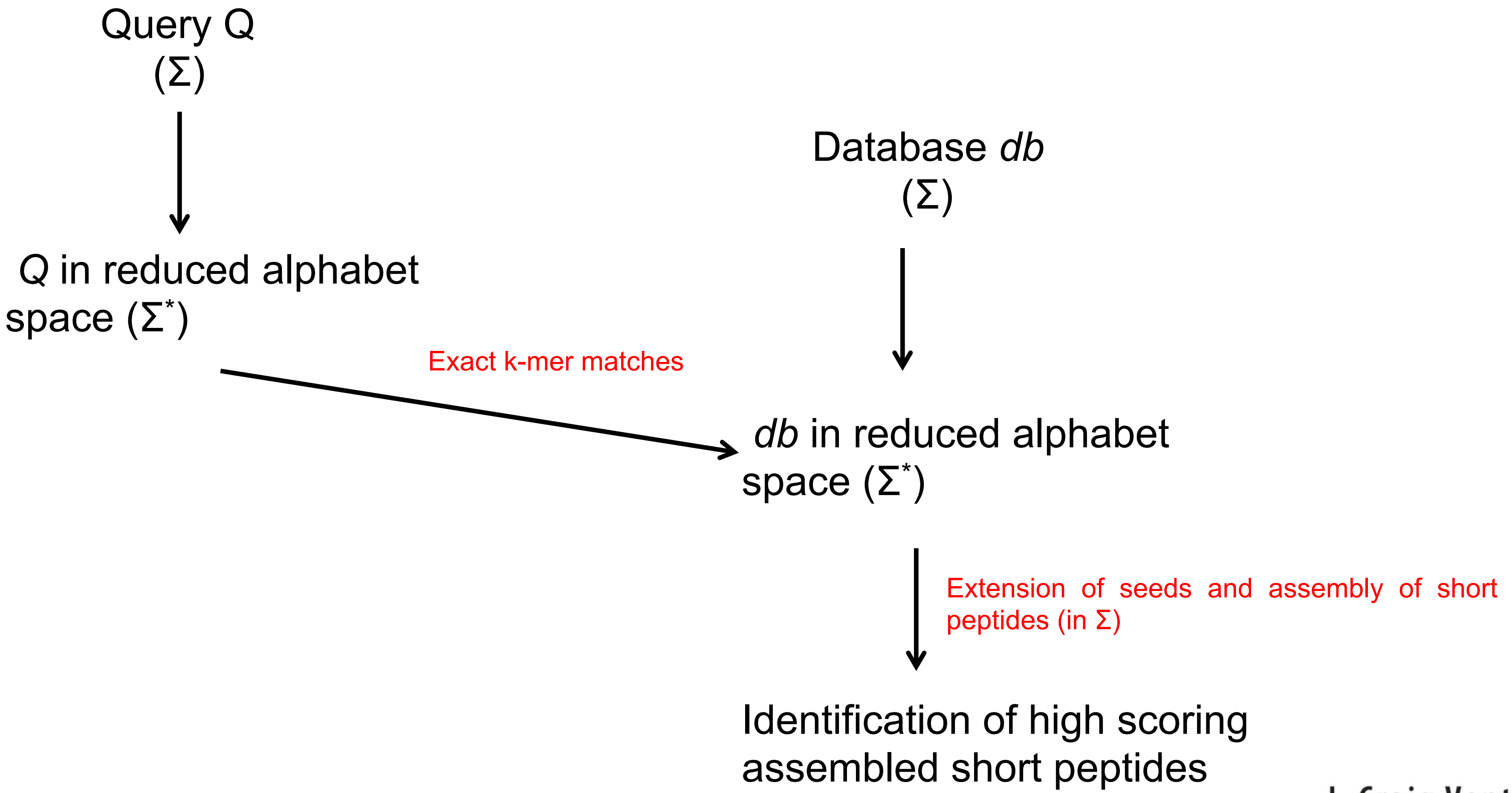
Construct a k-mer graph G of sequences in *db*



Need to check that the path has support of peptides in *db*

GRASP strategy

(Zhong, Yang, and Yooseph, in prep)



Query Q
(Σ)

Q in reduced alphabet space (Σ^*)

Exact k-mer matches

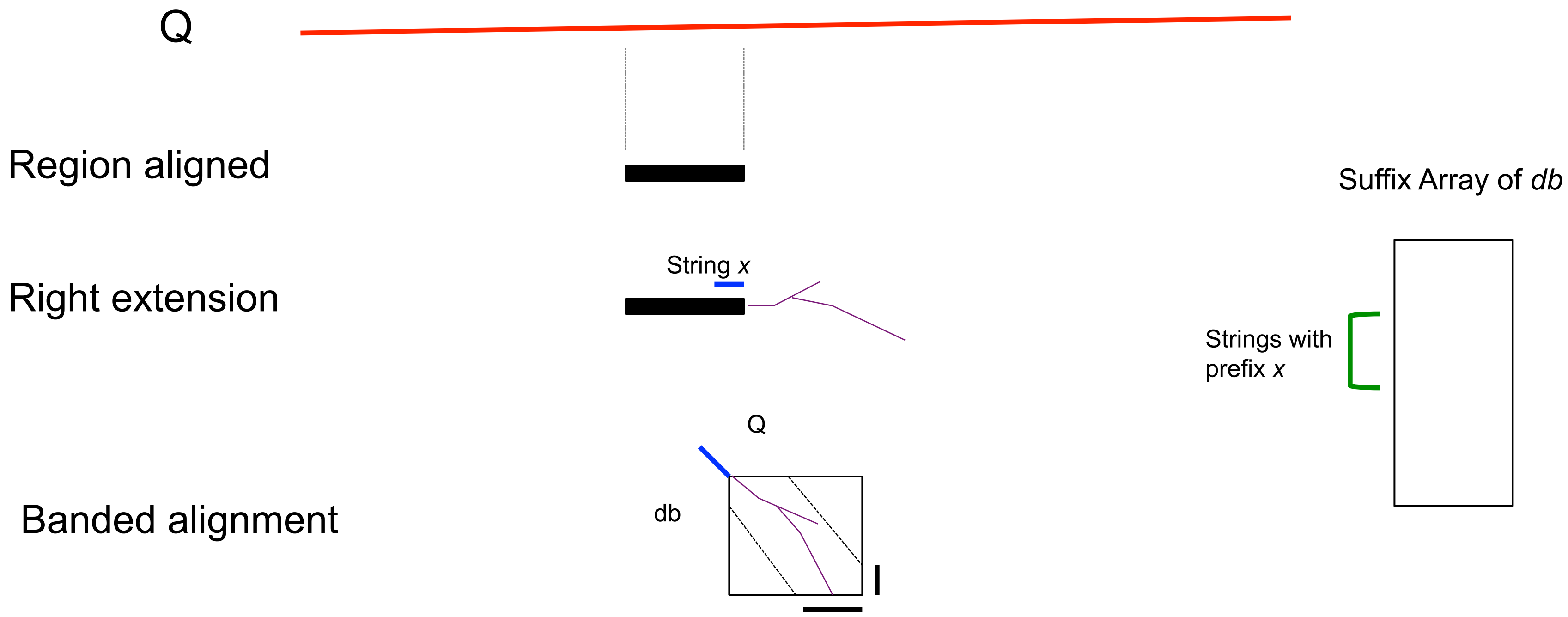
Database *db*
(Σ)

db in reduced alphabet space (Σ^*)

Extension of seeds and assembly of short peptides (in Σ)

Identification of high scoring assembled short peptides

GRASP – extension step



Q

Region aligned

Right extension

Banded alignment

Check for support and placement of short peptides

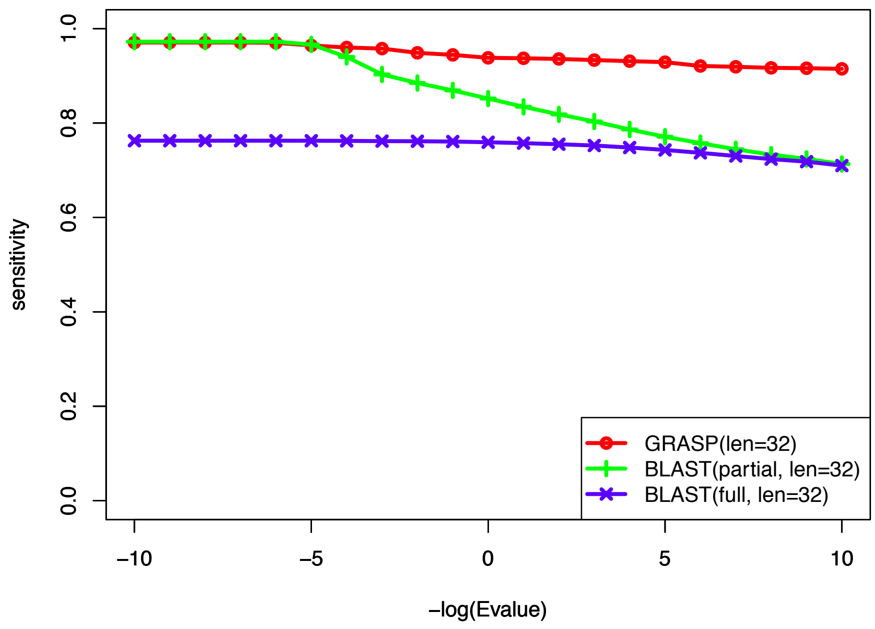
Left extension (similar fashion using reverse suffix array of *db*)

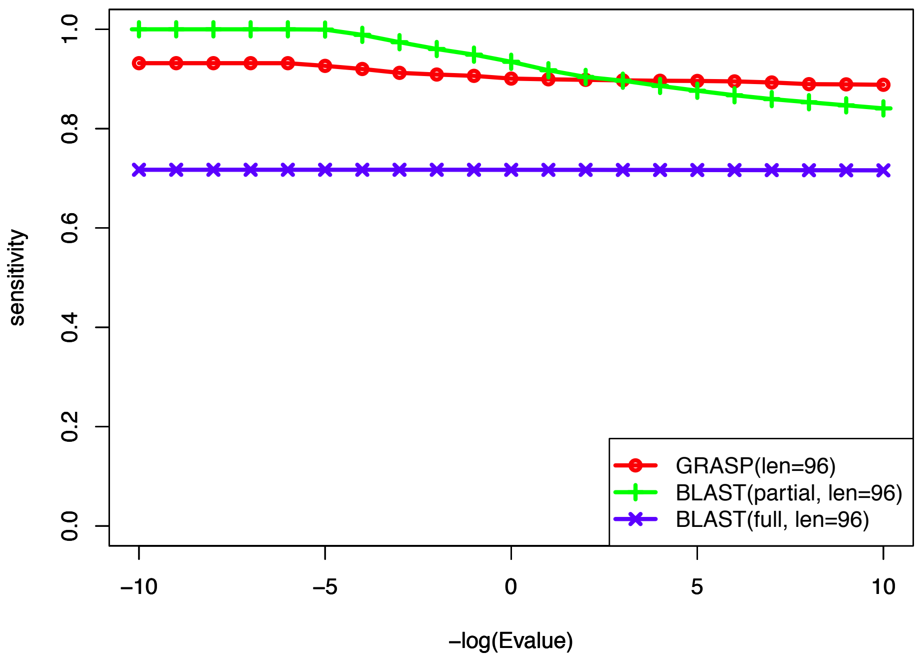
GRASP output

- Master-slave multiple sequence alignment of short peptides (slave) to query Q (master)

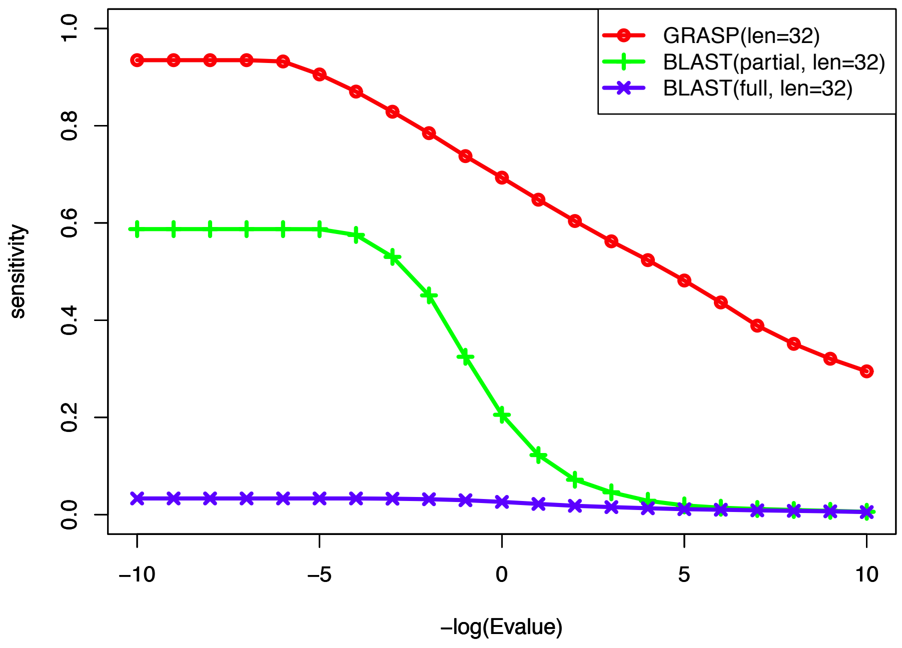
- Certificate
  - To convey evidence that a given short peptide, while individually may not meet the score (or E-value) threshold, does so when assembled with other short peptides
  - Composed of the assembly graph

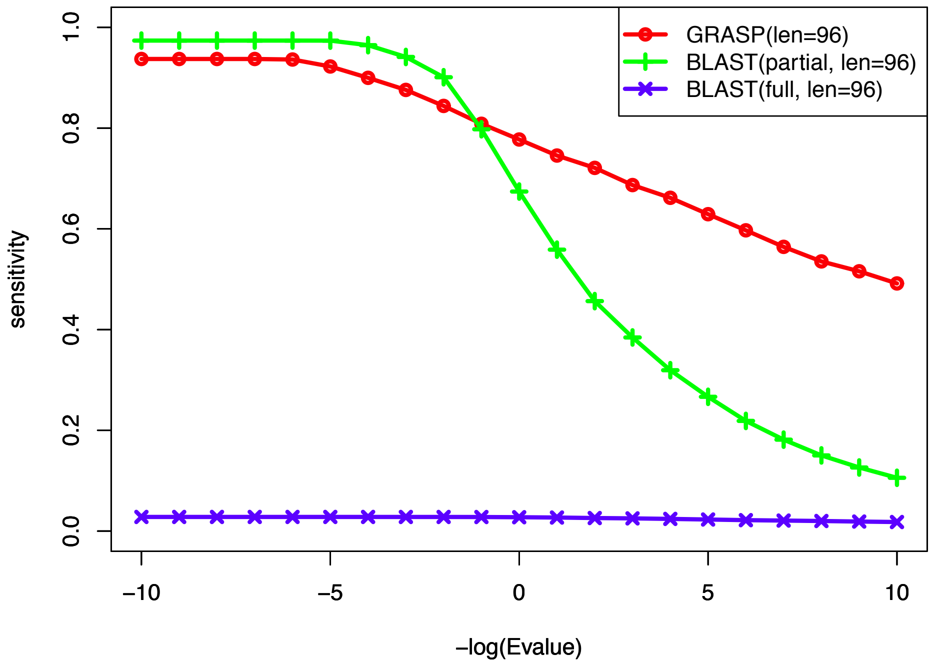
Example: RNA polymerase beta subunit (PF04563)



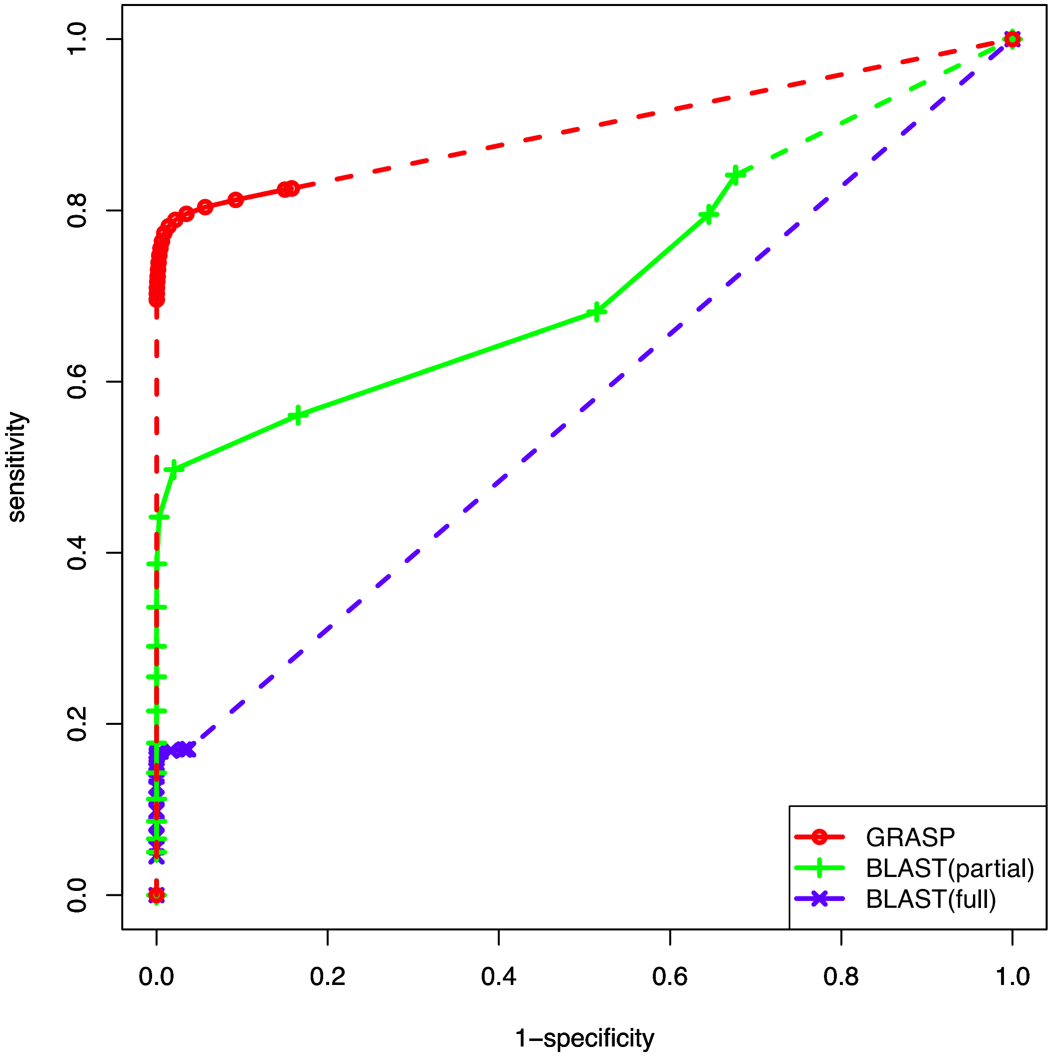


Example: LigT like Phosphoesterase (PF02834)





Glycosyl hydrolase superfamily members
(PF00128, PF00933, PF01120, PF12888, PF14701)



ROC curve

J. Craig Venter™ INSTITUTE

Conclusions

- SPA and GRASP are promising approaches for analyzing proteins in metagenomic datasets

- Resulting peptide assemblies could be used as a starting point for studying protein family evolution and function, and for inferring metabolic potential of constituent microbes

Acknowledgements

- Youngik Yang and Cuncong Zhong

- Funding
  - National Institutes of Health (AI084844, CA140233)
  - Beyster Family Foundation Fund of the San Diego Foundation
  - Life Technologies Foundation
  - Baltic Sea 2020 Foundation
  - Olle Enkvist Byggma Foundation
  - National Science Foundation (DBI-1262295 PI: Yooseph)

J. Craig Venter™ INSTITUTE