A Galaxy metaomic workflow for reference-tree based phylogenetic placement (MG-RTPP)

Ambrose Andongabo, Ian M. Clark, Dariush Rowlands, Keywan Hassani-Pak, Penny R. Hirsch, Elisa Loza-Reyes and **Andrew L. Neal**

BBSRC bioscience for the future

ROTHAMSTED RESEARCH

Systems biology for soil

encompassing soils' full complexity: biology, chemistry and physics in a dynamic structured habitat

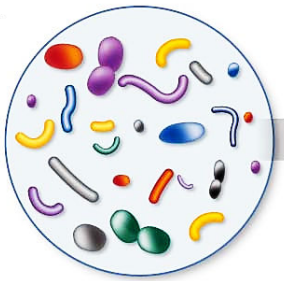
- community process rates
- gene phylogenetic diversity and expression,
- soil structure-biological function relationships,
- integration of theory and experiment

ROTHAMSTED RESEARCH

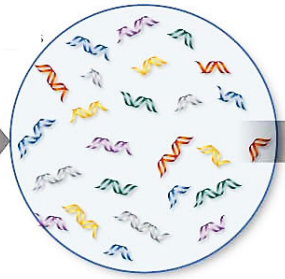
BBSRC bioscience for the future

The Challenge

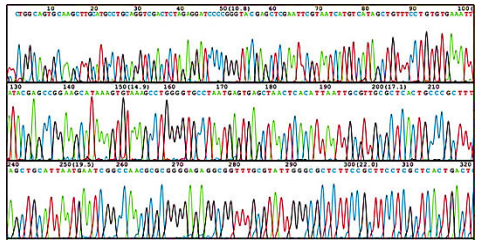
10^{12} (trillion) individual bacteria, the majority have never been cultured or studied



microbiome



extracted DNA or mRNA



high-throughput sequencing

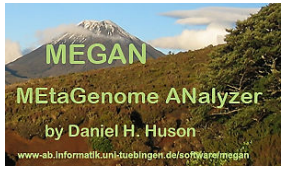
- Large datasets – *e.g.*~314 million reads of 150 nucleotide bases per sample
- Many genes of interest are poorly represented in databases and poorly studied

Quantitative Insights Into Microbial Ecology

qiime

taxon or species level analysis – no function

MG-RAST metagenomics analysis server



protein, functional level – no phylogeny

?

phylogenetic diversity of functional genes

Matsen et al. BMC Bioinformatics 2010, 11:538
http://www.biomedcentral.com/1471-2105/11/538

BMC Bioinformatics

Open Access

METHODOLOGY ARTICLE

pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree

Frederick A Matsen[1*], Robin B Kodner[2,3], E Virginia Armbrust[2]

Abstract

Background: Likelihood-based phylogenetic inference is generally considered to be the most reliable classification method for unknown sequences. However, traditional likelihood-based phylogenetic methods cannot be applied to large volumes of short reads from next-generation sequencing due to computational complexity issues and lack of phylogenetic signal. "Phylogenetic placement," where a reference tree is fixed and the unknown query sequences are placed onto the tree via a reference alignment, is a way to bring the inferential power offered by likelihood-based approaches to large data sets.

Results: This paper introduces `pplacer`, a software package for phylogenetic placement and subsequent visualization. The algorithm can place twenty thousand short reads on a reference tree of one thousand taxa per hour per processor, has essentially linear time and memory complexity in the number of reference taxa, and is easy to run in parallel. `Pplacer` features calculation of the posterior probability of a placement on an edge, which is a statistically rigorous way of quantifying uncertainty on an edge-by-edge basis. It also can inform the user of the positional uncertainty for query sequences by calculating expected distance between placement locations, which is crucial in the estimation of uncertainty with a well-sampled reference tree. The software provides visualizations using branch thickness and color to represent number of placements and their uncertainty. A simulation study using reads generated from 631 COG alignments shows a high level of accuracy for phylogenetic placement over a wide range of alignment diversity, and the power of edge uncertainty estimates to measure placement confidence.

Conclusions: `Pplacer` enables efficient phylogenetic placement and subsequent visualization, making likelihood-based phylogenetics methodology practical for large collections of reads; it is freely available as source code, binaries, and a web service.

Background

High-throughput pyrosequencing technologies have enabled the widespread use of metagenomics and metatranscriptomics in a variety of fields [1]. This technology has revolutionized the possibilities for unbiased surveys of environmental microbial diversity, ranging from the human gut to the open ocean [2-8]. The trade off for high throughput sequencing is that the resulting sequence reads can be short and come without information on organismal origin or read location within a genome.

The most common way of analyzing a metagenomic data set is to use BLAST [9] to assign a taxonomic name to each query sequence based on "reference" data of known origin. This strategy has its problems: when a query sequence is only distantly related to sequences in the database, BLAST can either err substantially by forcing a query into an alignment with a known sequence, or return an uninformatively broad collection of alignments. Furthermore, similarity statistics such as BLAST *E*-values can be difficult to interpret because they are dependent on fragment length and database size.

* Correspondence: matsen@fhcrc.org
[1]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
Full list of author information is available at the end of the article

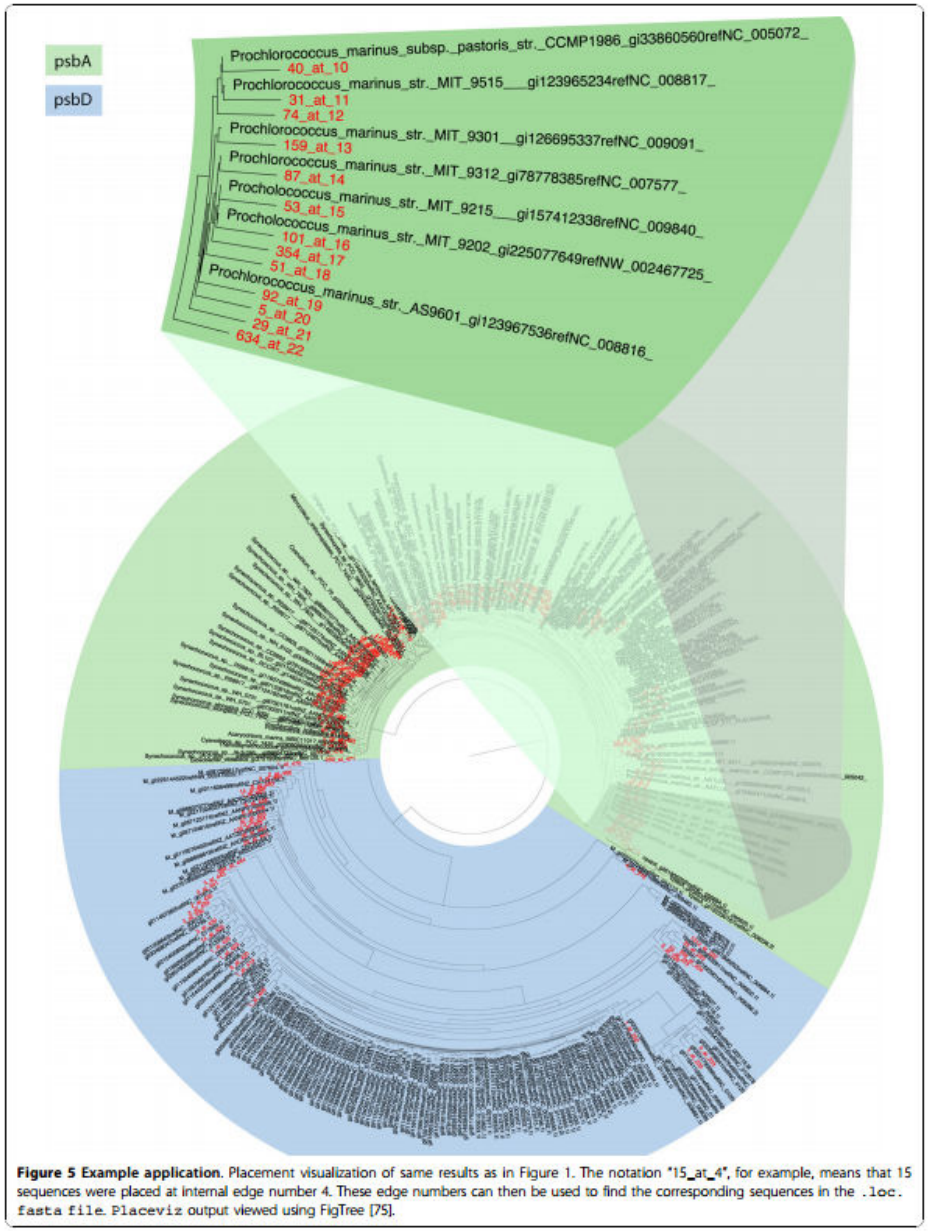
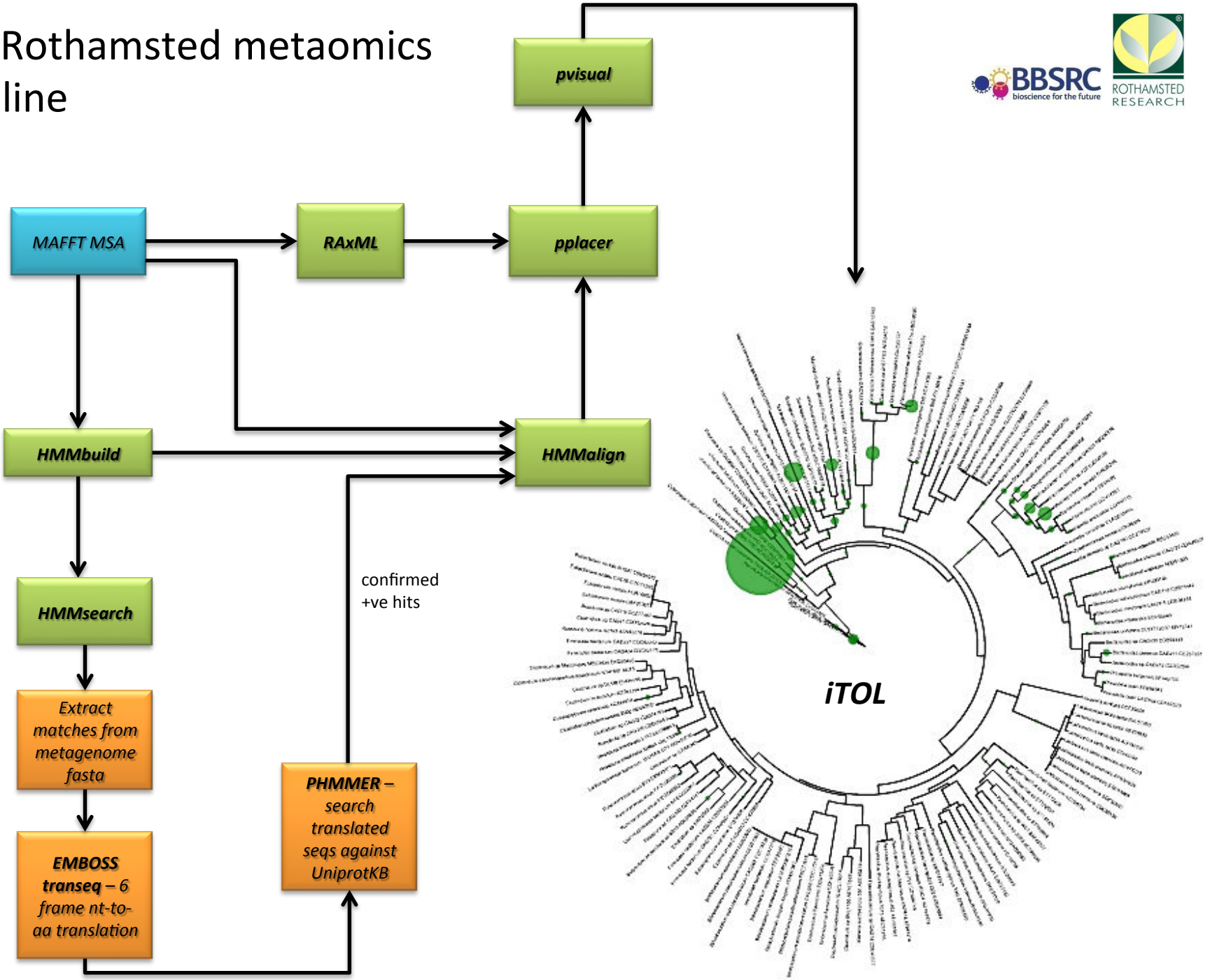


Figure 5 Example application. Placement visualization of same results as in Figure 1. The notation "15_at_4", for example, means that 15 sequences were placed at internal edge number 4. These edge numbers can then be used to find the corresponding sequences in the `.loc. fasta file`. `Placeviz` output viewed using FigTree [75].

Matsen et al., BMC Bioinformatics 2010, 11: 538

The Rothamsted metaomics pipeline



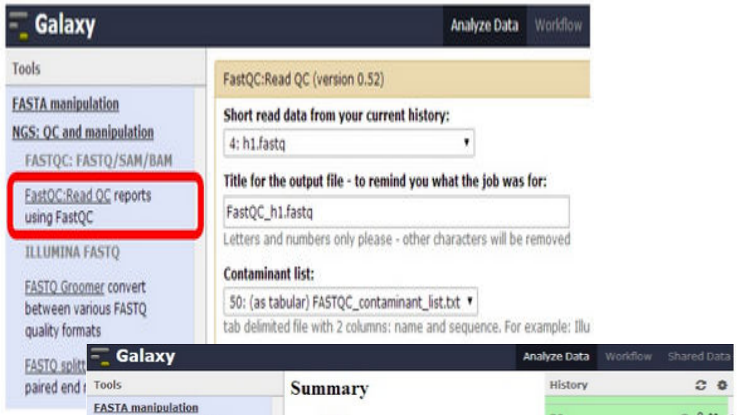
Main components of the pipeline

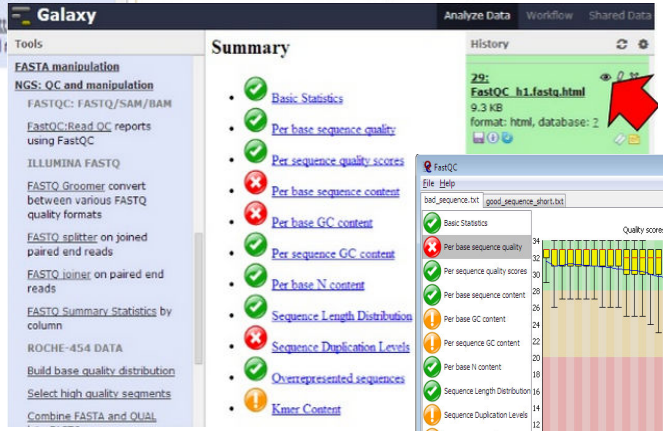
ROTHAMSTED RESEARCH

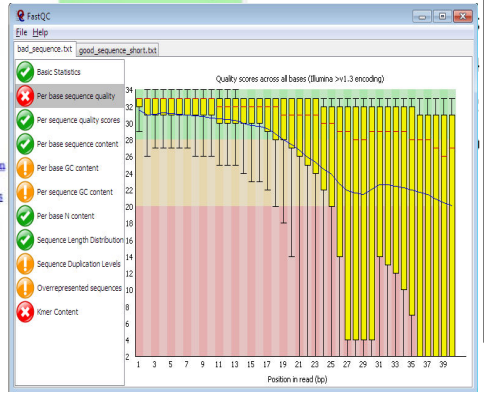
- Pre-processing the ‘metaomics’ dataset
- Generating the reference dataset
- HMMER tools used within the pipeline
- Confirming positive hits
- Sections for data extraction, data conversion and data visualization

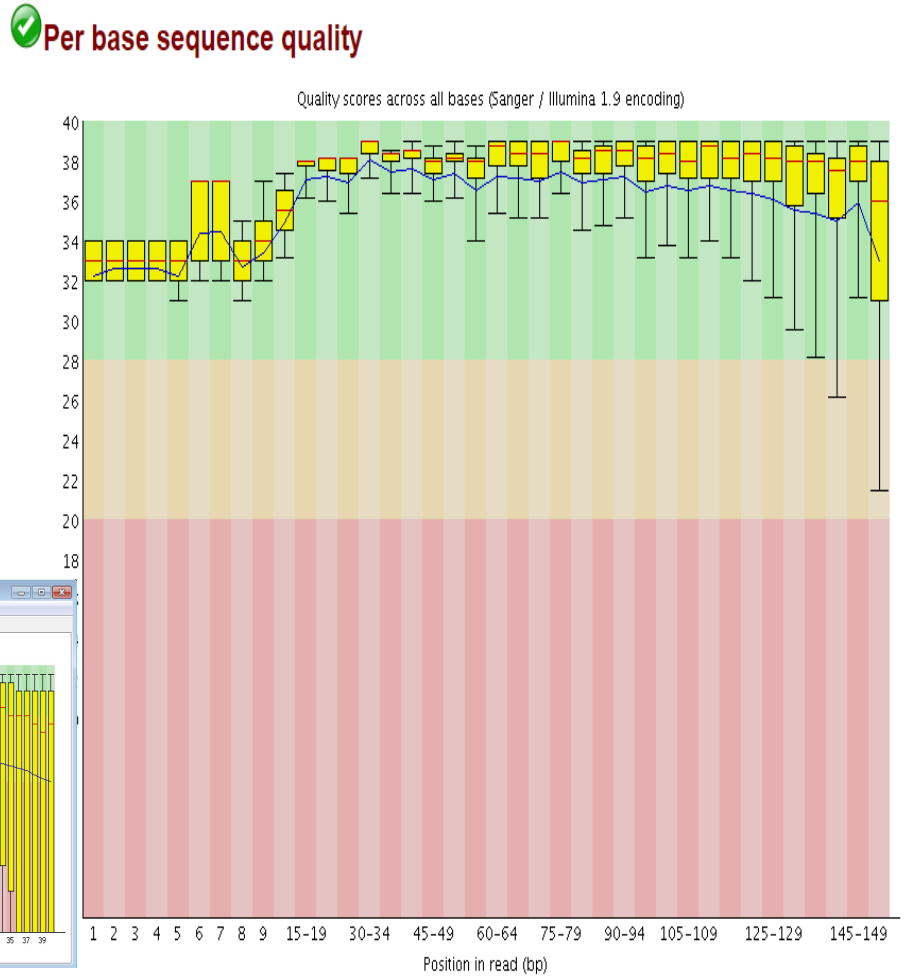
BBSRC bioscience for the future

Processing the ‘metaomics’ file with quality dataset









Generating the reference Multiple Sequence Alignment (MSA)

ROTHAMSTED RESEARCH

1) Self generated MSA or MSA generated by others

2) Assemble a collection of full-length genes sequences

a) Start with a small collection of well characterized proteins

b) Use this to generate large collections using BlastP or use jackhmmer to search for homologous proteins in UniprotKB

c) A larger collection of protein is then generated based upon a common pHMM. It is important to check it to remove partial sequences or obvious outliers

d) Extract corresponding nucleotide sequences from ENA and then use any tool to generate MSA (e.g MAFFT)

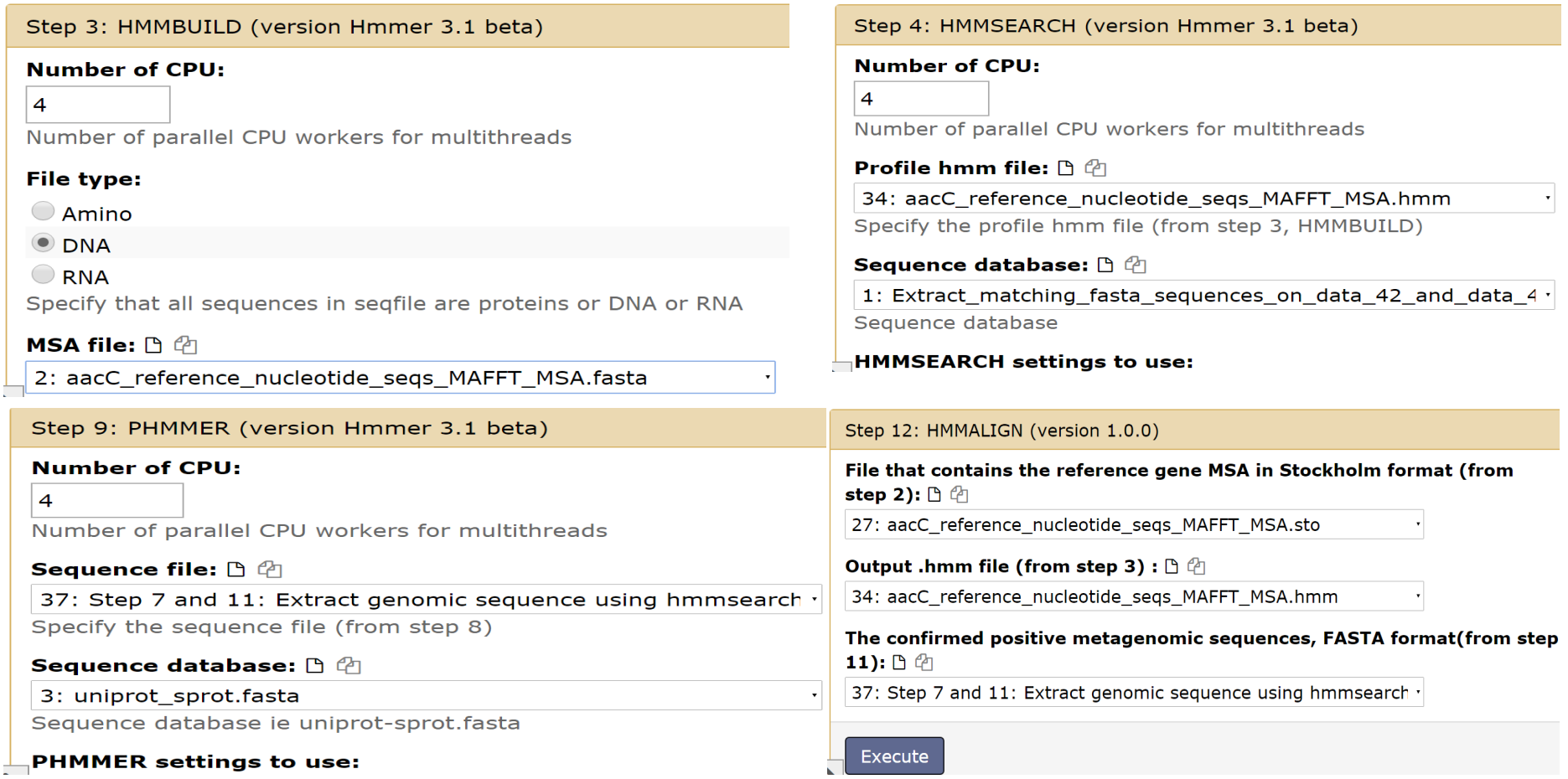
3) Another option is to generate a reference set of gene sequences from the SEED database.

BBSRC bioscience for the future

The core hmmer tools used within the pipeline

ROTHAMSTED RESEARCH

Hmmbuild, **Hmmsearch**, Hmmscan, **Hmmalign**, **Phmmer**, Nhmmer , Hmmconvert Hmmemit, Hmmfetch, Hmmpress, Hmmsim, Hmmstat and **Jackhmmer**



BBSRC bioscience for the future

Detailed view of HMMSEARCH Wrapper

ROTHAMSTED RESEARCH

HMMSEARCH settings to use:

Full parameter list

You can use the default settings or set custom values for any of hmmsearch's parameters.

Do you want to control reporting and inclusion thresholds:

Yes

Select which threshold option to use:

Show target sequences hits and domain hits that meet the chosen REPORTING bit scores threshold

- Show target sequences hits and domain hits that meet the chosen REPORTING E-values threshold
- Show target sequences hits and domain hits that meet the chosen REPORTING bit scores threshold
- Show target sequences hits and domain hits that meet the chosen INCLUSION E-values threshold
- Show target sequences and domain hits that meet the chosen INCLUSION bit scores threshold
- Show target sequences and domain hits that meet both the chosen REPORTING and INCLUSION E-value threshold
- Show target sequences and domain hits that meet both the chosen REPORTING and INCLUSION bit scores threshold

e thresholding

Reporting domain bit score threshold:

800.0

Report domains having bit scores less than or equal to this bit score threshold in output

Do you want to control the acceleration pipeline:

Yes

Do you want to control acceleration heuristicss:

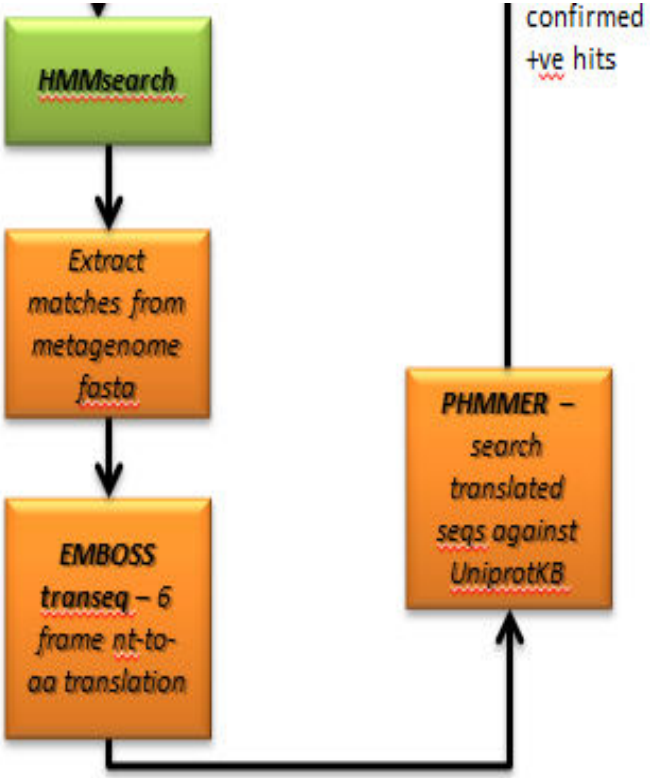
Turn off all filters, including the bias filter

Options controlling acceleration heuristics

Turn all heuristic filters off (less speed, more power):

BBSRC bioscience for the future

Workflow section used to confirm the positive hits



Why confirm the hits?

To be absolutely sure that the results from the hmmer search is a set with no false positives

How to confirm the hits

Perform transeq translation of the hmmsearch results and search the translated sequences against UniprotKB

Tool wrappers

ROTHAMSTED RESEARCH

Tools

Step 1: Fasta2Phy - Convert reference MSA in FASTA format to relaxed Phylip

Step 2: Phy_to_Sto converting Phylip to Stockholm format

Step 3: HMMBUILD - profile HMM construction from multiple sequence alignments

Step 5: RaXML Maximum Likelihood based inference of large phylogenetic trees

Step 4: HMMSEARCH - search profile(s) against a sequence database to generate sequence IDs and homology scores table

Step 7 and 11: Extract genomic sequence using hmmsearch results for each target id in the file

Step 6: Extract sequence Ids from an HMMSearch result file

Step 8: SeqTranslator - 6 frame amino acid translator from genomic sequences. Generates 1 file per frame

Step 10: PhmmerGetSeqIDs Extracting metagenomic reads ids from phmmer tabular output

Step 9: PHMMER Confirmation step. Search protein sequence (ie 1-6 frame amino acid translations) against protein database

Step 12: HMMALIGN - align sequences to a HMM profile using an alignment filem its HMM profile and metagenome or metatranscriptome positively confirmed sequences

Wrappers for data conversion , data extraction and data visualization

ROTHAMSTED RESEARCH

DIRECT EXTRACTION OF SEQUENCES FROM ENA

Get the sequences from ENA using protein ACC/ID for a list of protein IDs in a file

REFSEQ: HAVING REFSEQ PROTEIN IDS START HERE

Refseq_Uniprot Mapping Refseq IDs to Uniprot ID for each refseq protein ID in a file

Step 13: Pplacer Place query sequences on a fixed reference phylogenetic tree to maximize phylogenetic likelihood or posterior probability according to a reference

Step 14: Pvisual Placement Visualization of reads on a phylogenetic tree

Step 1: Fasta2Phy - Convert reference MSA in FASTA format to relaxed Phylip

Step 2: Phy_to_Sto converting Phylip to Stockholm format

Step 7 and 11: Extract genomic sequence using hmmsearch results for each target id in the file

Step 6: Extract sequence Ids from an HMMSearch result file

Step 8: SeqTranslator - 6 frame amino acid translator from genomic sequences. Generates 1 file per frame

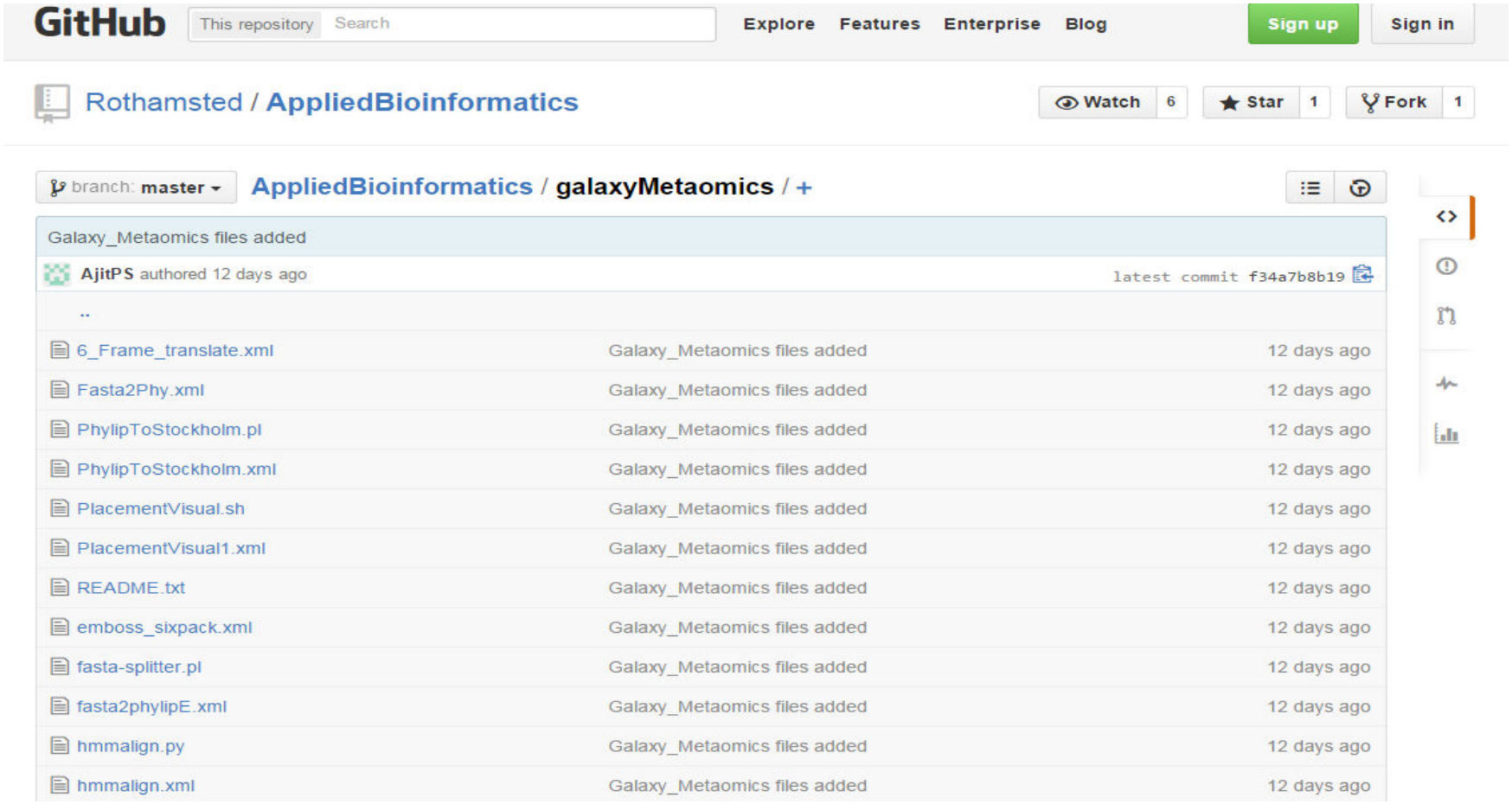
BBSRC bioscience for the future

Metaomics pipeline in our galaxy instance

ROTHAMSTED RESEARCH

BBSRC bioscience for the future

Where can you find the tools and the tools wrappers



https://github.com/Rothamsted/AppliedBioinformatics/tree/master/galaxyMetaomics

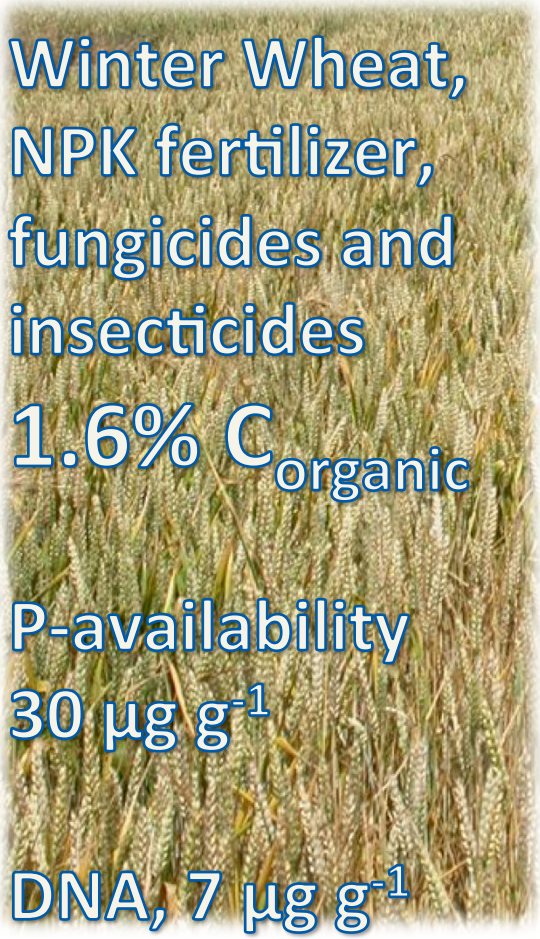
We are hiring Bioinformatics Scientist

BBSRC bioscience for the future

ROTHAMSTED RESEARCH

Example application: land-use effects upon phosphorus cycle genes

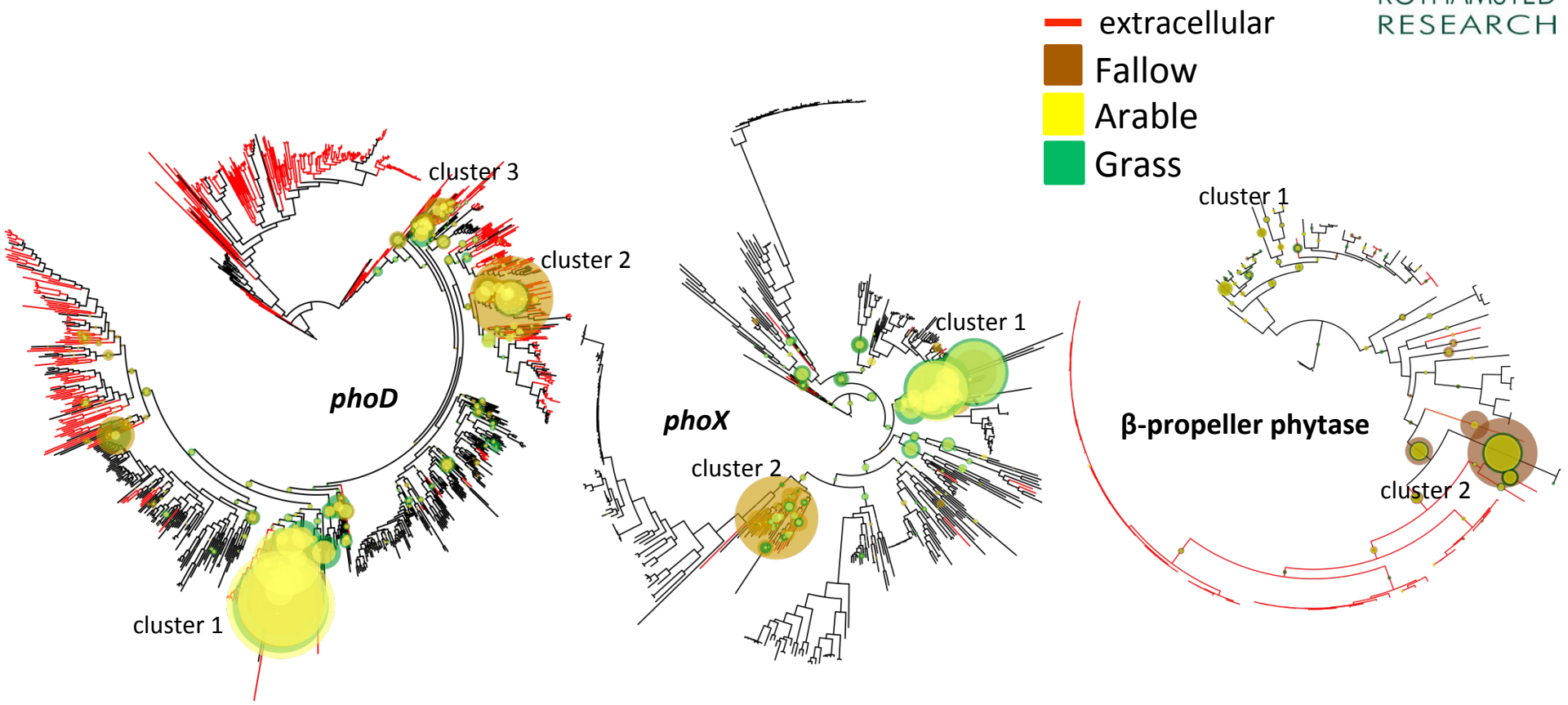
ROTHAMSTED RESEARCH



Hirsch *et al*., 2009 *Soil Biol Biochem* 41, 2021-2024.

BBSRC bioscience for the future

Results



Trees presented in *iTOL* allowing multiple layers of information – phylogenetic structure, subcellular localisation and multiple datasets

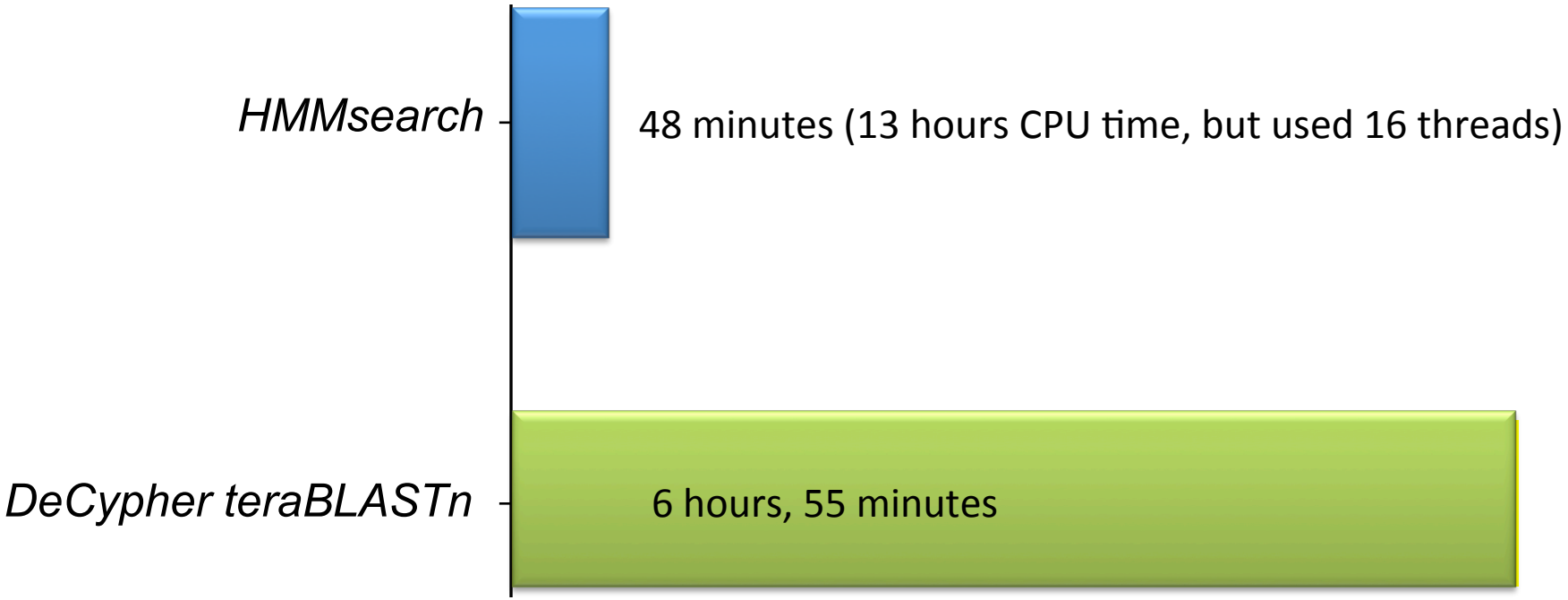
Acknowledgement

Keywan Hassani-Pak
Andrew L. Neal
Dariush Rowlands
Ian M. Clark
Chris Rawlings
Penny R. Hirsch
Eliza Loza-Reyes

Sean R. Eddy
Travis J. Wheeler
Robert D. Finn
Jody Clements
William Arndt
Benjamin L. Miller
Fabian Schreiber
Alex Bateman

pHMM *versus* BLAST: how do the two approaches compare?

To search for 976 reference nucleotide sequences (*rpoB*, 4,029 bases) in a 576,395,603 sequence (100 bases) dataset……



pHMM *versus* BLAST: how do the two approaches compare?