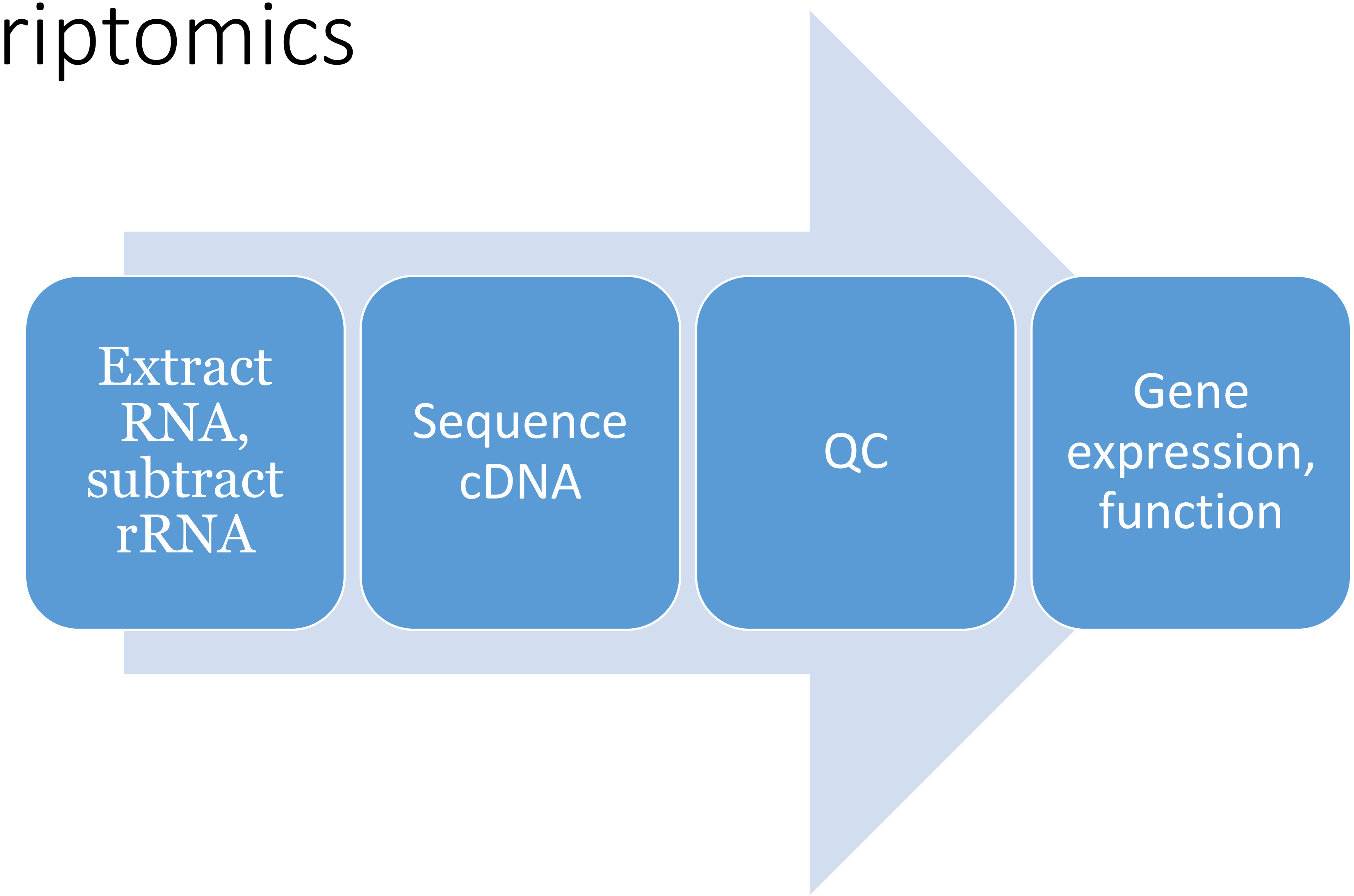
CBC Data Therapy

Metatranscriptomics Discussion

Computational Biology Core

UCONN
UNIVERSITY OF CONNECTICUT

Metatranscriptomics



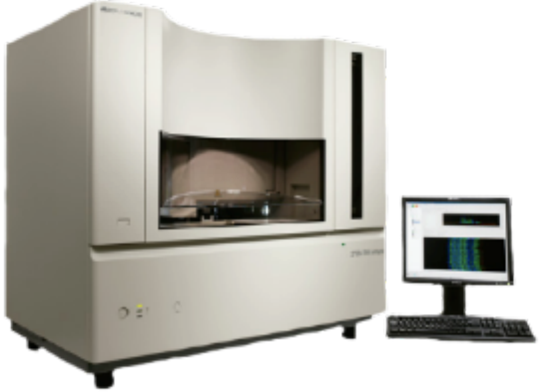
Extract RNA, subtract rRNA

Sequence cDNA

QC

Gene expression, function

Sequencing

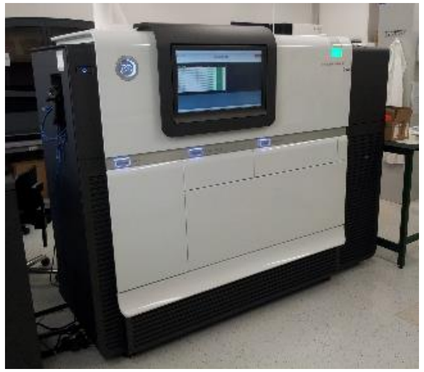


Sanger

Ion Torrent

Roche 454

Illumina *Seq



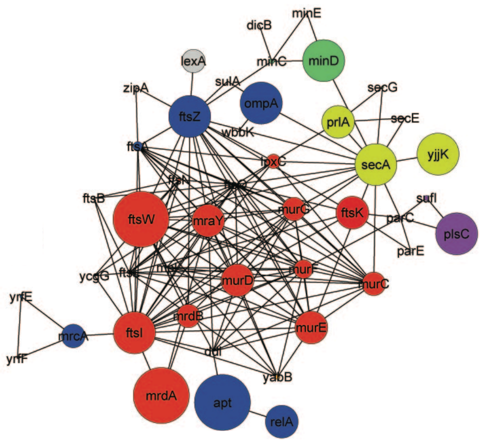
Pacific Biosciences

Nanopore

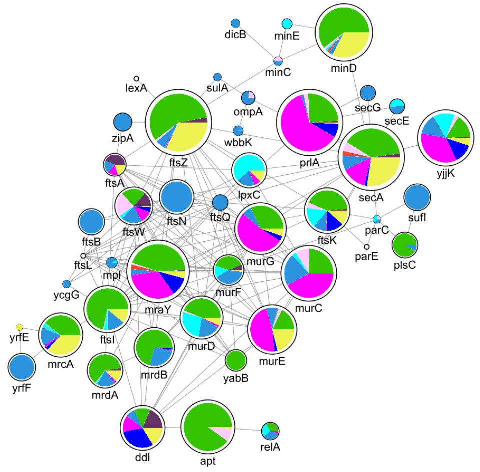
Metatranscriptomics focuses on community activity

Metatranscriptomics exploits RNA-Seq to determine which genes and pathways are being actively expressed within a community

Genes involved in pathways associated with cell wall biogenesis



Relative abundance

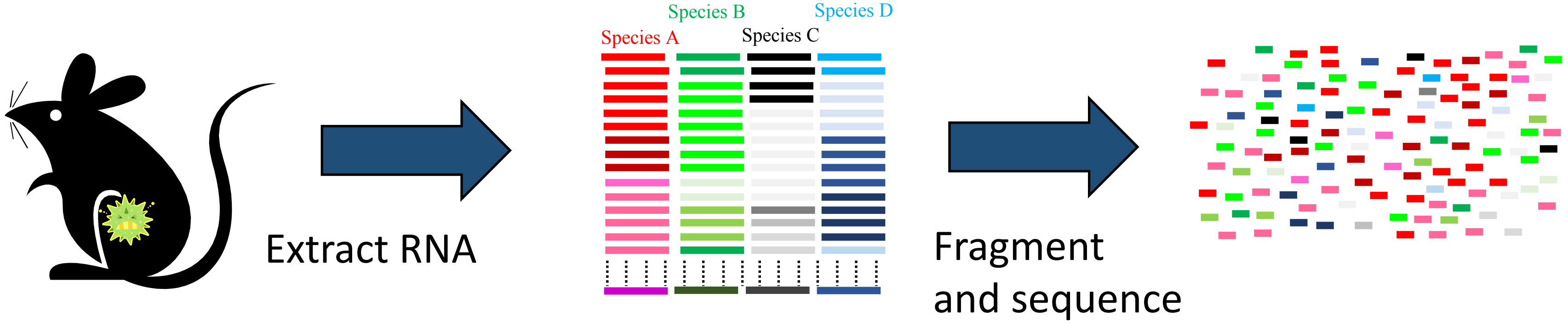


Relative abundance and contributing taxa

Metatranscriptomics can reveal active *functions* (knowing the taxa responsible is unimportant)

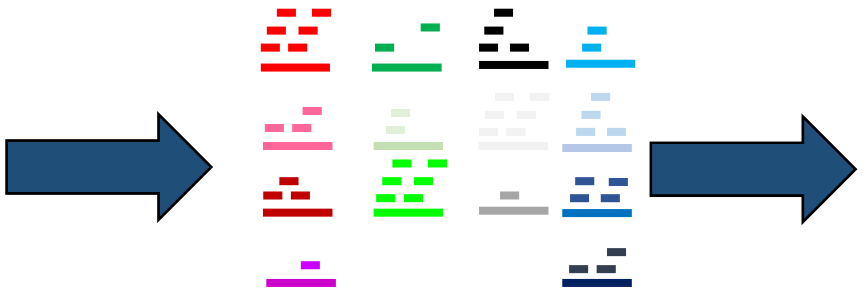
It can also reveal which taxa are responsible for the active functions

Metatranscriptomics through RNA Seq



Extract RNA

Fragment and sequence



Align reads to known transcripts to obtain relative expression

Gene A1	6
Gene A2	3
Gene A3	3
Gene A4	1
Gene B1	2
Gene B2	2
Gene B3	6
Gene C1	4
⋮	⋮
Gene D4	3

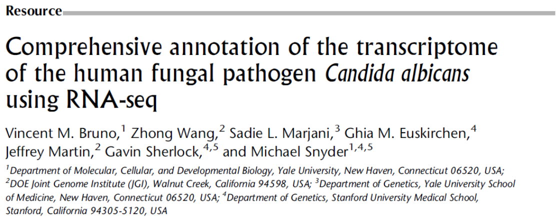
RNA-Seq is the unbiased sequencing of an RNA sample to yield a digital readout of the relative expression of transcripts within a sample

Typically applied to organisms with a reference (sequenced) genome, microbiome applications face a number of challenges

III IV

Metatranscriptomics: Challenges

In a typical RNA-Seq experiment applied to a single eukaryotic organism, mRNA is isolated. After fragmentation and sequencing, reads are mapped to a reference genome using standard software such as MAQ and BWA to provide: 1) support that the transcript is expressed; 2) the relative abundance of the transcript; and 3) the presence and abundance of isoforms

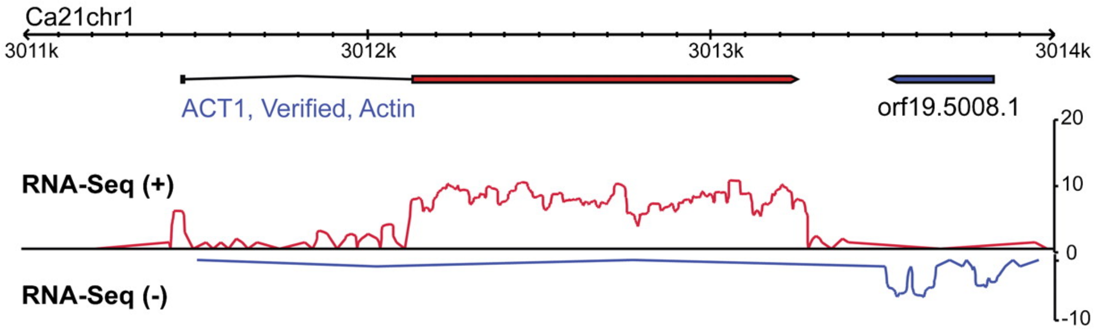


Resource

Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq

Vincent M. Bruno,[1] Zhong Wang,[2] Sadie L. Marjani,[3] Ghia M. Euskirchen,[4] Jeffrey Martin,[2] Gavin Sherlock,[4,5] and Michael Snyder[1,4,5]

[1]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA; [2]DOE Joint Genome Institute (JGI), Walnut Creek, California 94598, USA; [3]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [4]Department of Genetics, Stanford University Medical School, Stanford, California 94305-5120, USA

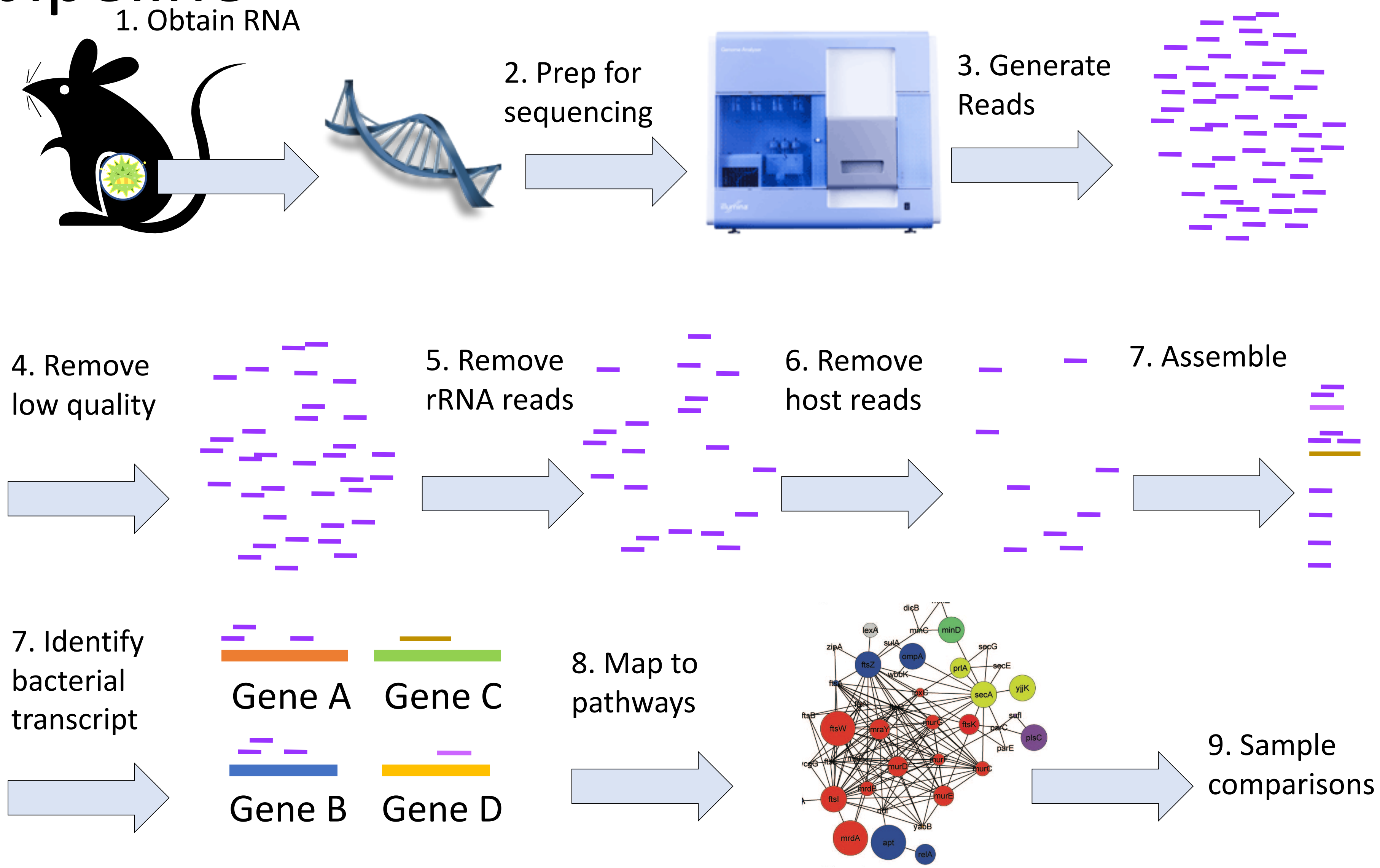


For microbiome samples we have the following problems:

a) Lack of a polyA signal makes it difficult to isolate bacterial mRNA and resulting in (massive) rRNA contamination
b) Environmental microbiome samples lack reference genomes making it difficult to map reads back to their source transcripts

III IV

A typical metatranscriptomic analytical pipeline



1. Obtain RNA

2. Prep for sequencing

3. Generate Reads

4. Remove low quality

5. Remove rRNA reads

6. Remove host reads

7. Assemble

7. Identify bacterial transcript

Gene A

Gene C

Gene B

Gene D

8. Map to pathways

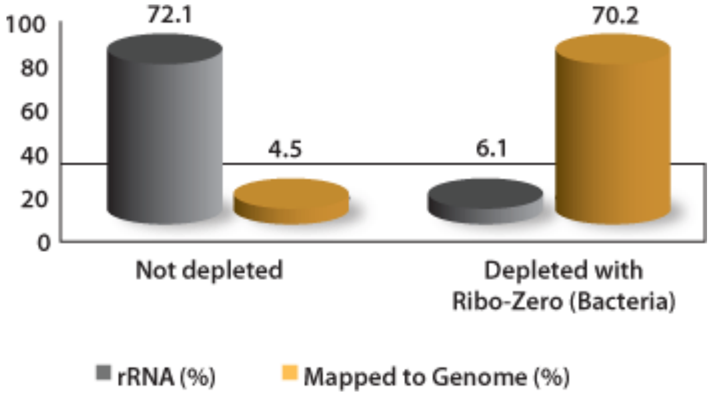
9. Sample comparisons

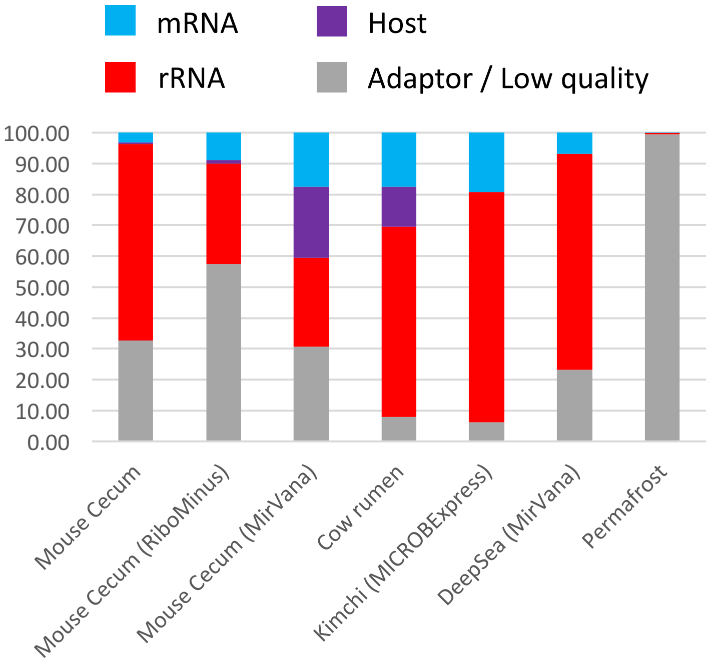
Preparing sample for sequencing

Bacterial mRNA’s lack a polyA tail so how to remove abundant rRNA species?

Once RNA has been extracted, several kits are available to remove rRNA – need 500ng-2.5ug RNA/sample

Ribo-Zero (Illumina) provides reasonable success

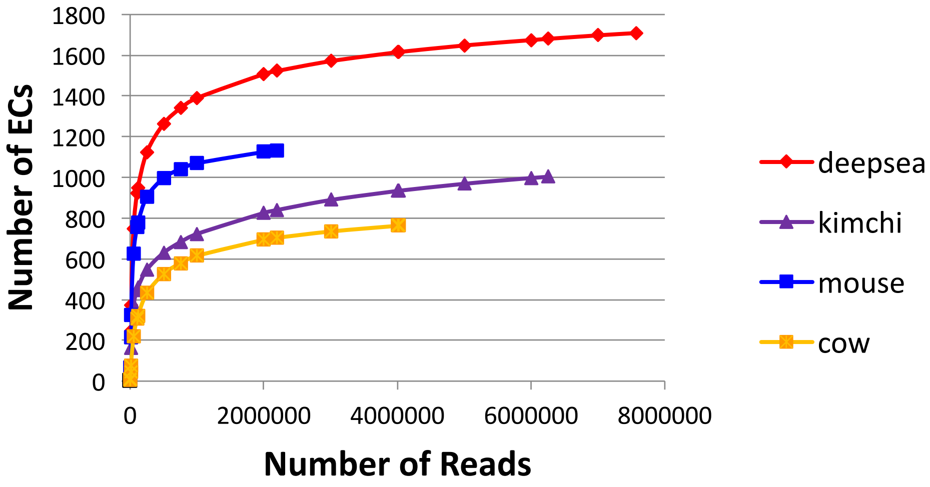




Host mRNAs can also prove challenging – can also be informative!

Generating reads

How many reads are “enough”?



20 million/sample mRNA

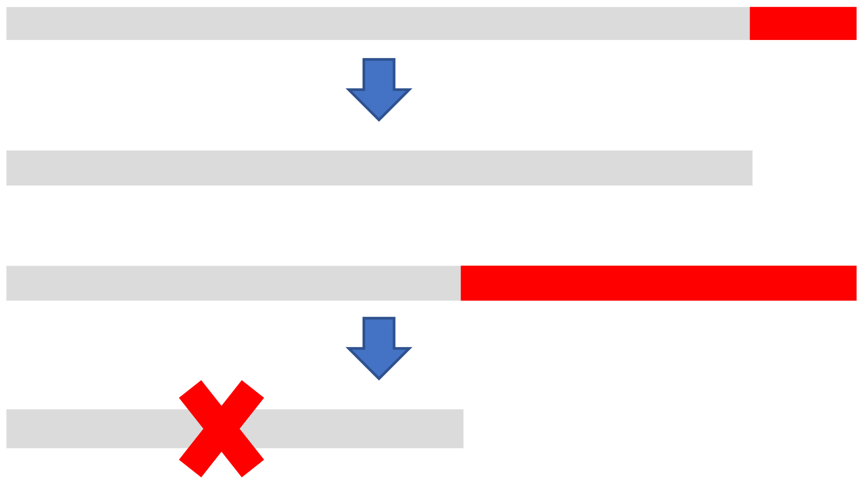
While PB and MiSeq provide long reads useful for annotation HiSeq (or NextSeq) provide sequencing depth and offer possibility of multiplexing

Read processing - filtering

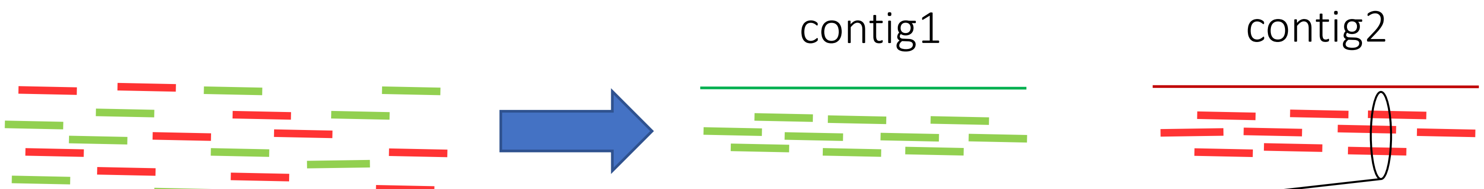
To identify reads derived from mRNA bioinformatics pipelines need to be in place that remove contaminating reads:

Low quality - *Trimmomatic*
Adaptors – *Trimmomatic* & *Cross_Match*
Host - *BWA*
rRNA – *BLAT / Infernal*

Trimmomatic uses a sliding window approach from the 5` end to identify low quality regions which are then trimmed from the 3` end. Reads < 36 bp are discarded

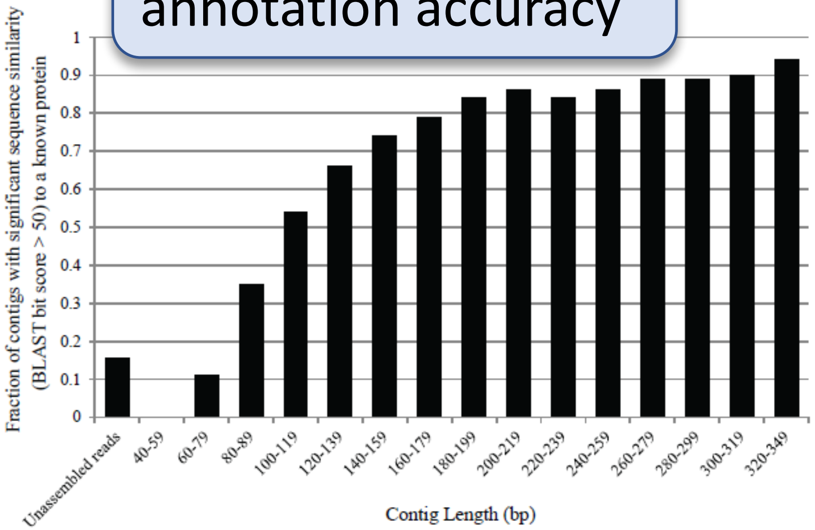


Read processing - Assembly



attagcggcgattttcggcgatcttatcttgatctgggcgcgtatcggtagcgtagcgattcgtagc
attagcggcgattttcggcgatcttatcttgatctgggcgcgtatcggtagcgtagcgattcgtagc
attagcggcgattttcggcgatcttatcttgatctgggcgcgtatcggtagcgtagcgattcgtagc

Assembly improves annotation accuracy



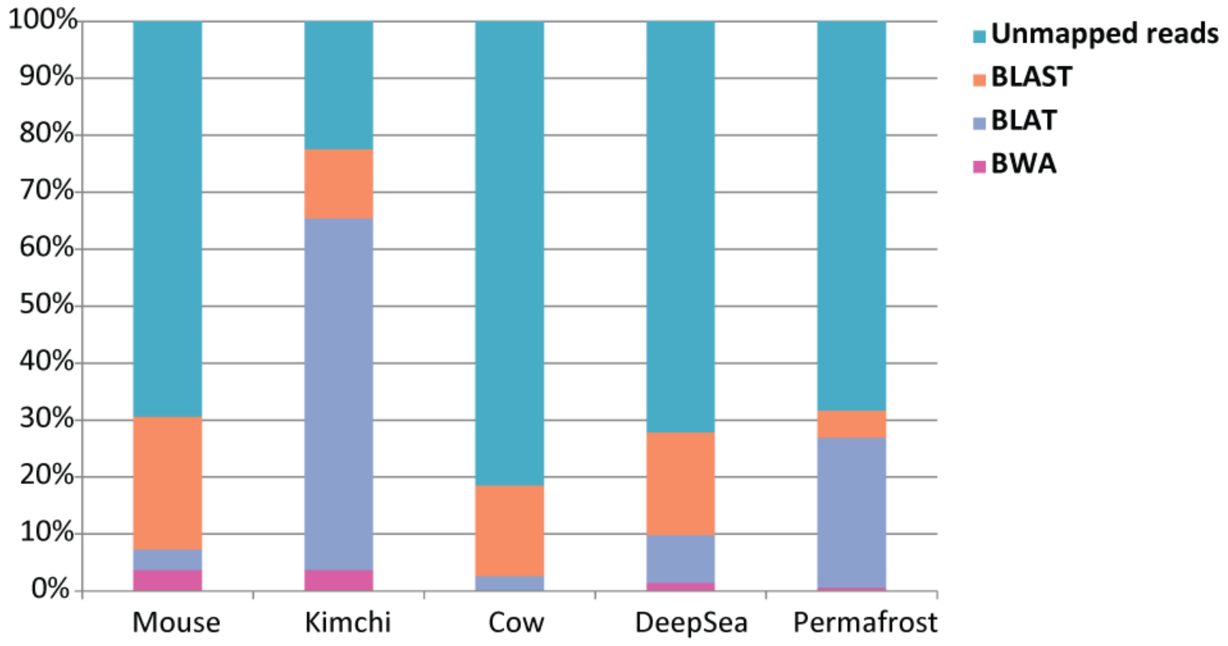
Trinity appears to provide best performance in terms of reads that can be annotated

Chimera’s, misassembled contigs, can become a problem due to reads derived from orthologs from different species

Read processing – functional annotation

One solution is to work in peptide space and use *BLASTX* to search protein databases - this is very time consuming and requires cloud/cluster computing

Other solutions USEARCH/VSEARCH or DIAMOND (issues over quality and cost)



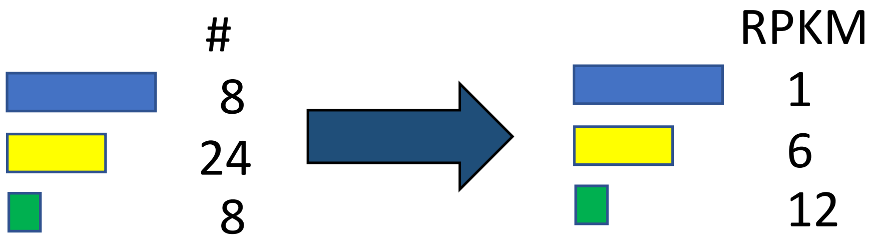
Even with BLAST many reads remain unannotated

Can be improved with longer read length

Read processing – converting mappings to expression

To normalize expression levels to account for differences in gene length, read counts are converted to ***Reads per kilobase of transcript mapped (RPKM)***

Expression is biased for gene length (longer transcripts should have more reads) to normalize, reads are converted to Reads per Kilobase of transcript per million reads mapped



$\mathrm{RPKM}_{\mathrm{geneA}} = 10^9\, C_{\mathrm{geneA}} / NL$

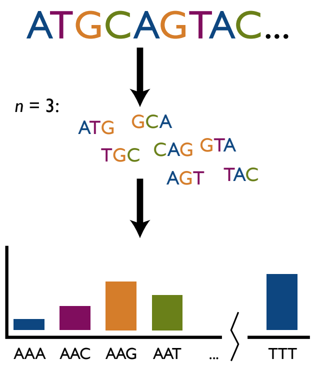
C_{geneA} = number of reads mapped to geneA
N = total number of reads
L = length of transcript in units of Kb

Several software tools available to do mapping and calculate normalized expression measurements across different samples including Bowtie and Cufflinks

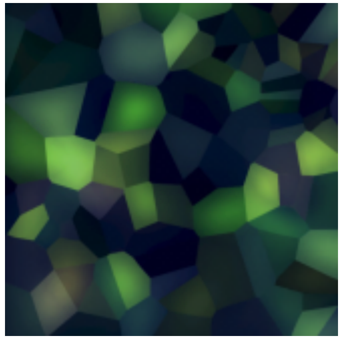
Read processing – taxonomic annotation

Alignment based methods such as BLAST and BWA can fail where we lack suitable reference genomes – particularly for short read datasets where assignments may be ambiguous

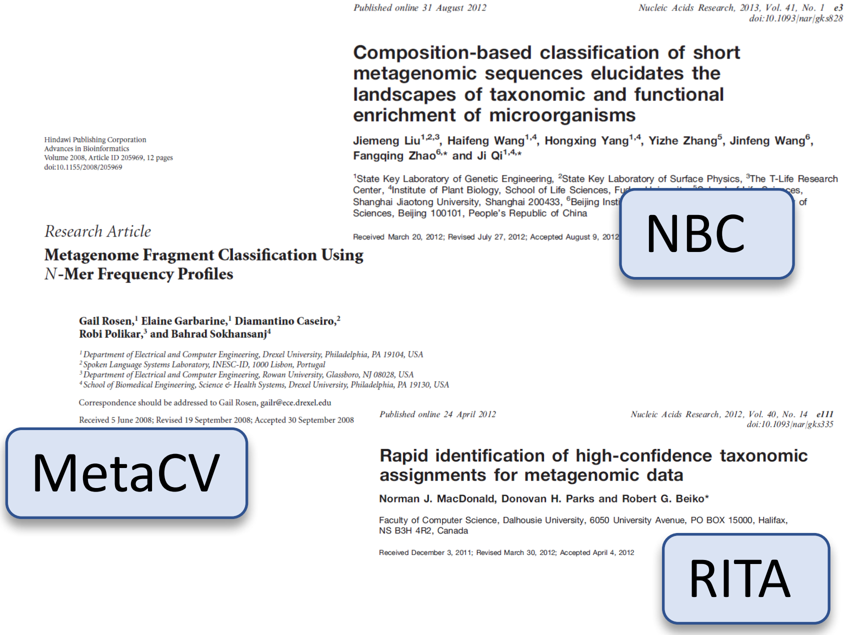
Compositional methods (e.g. nt frequency, codon bias) offer alternative strategies



Here a sequences is classified into frequencies of 3-mers



Nearest neighbours methods then try to assign a sequence to the genome with the closest distribution



Published online 31 August 2012 Nucleic Acids Research, 2013, Vol. 41, No. 1 **e3**
doi:10.1093/nar/gks828

Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms

Jiemeng Liu[1,2,3], Haifeng Wang[1,4], Hongxing Yang[1,4], Yizhe Zhang[5], Jinfeng Wang[6], Fangqing Zhao[6,*] and Ji Qi[1,4,*]

[1]State Key Laboratory of Genetic Engineering, [2]State Key Laboratory of Surface Physics, [3]The T-Life Research Center, [4]Institute of Plant Biology, School of Life Sciences, Fu ces, Shanghai Jiaotong University, Shanghai 200433, [6]Beijing Insti of Sciences, Beijing 100101, People's Republic of China

Received March 20, 2012; Revised July 27, 2012; Accepted August 9, 2012

NBC

Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2008, Article ID 205969, 12 pages
doi:10.1155/2008/205969

Research Article

Metagenome Fragment Classification Using *N*-Mer Frequency Profiles

Gail Rosen,[1] Elaine Garbarine,[1] Diamantino Caseiro,[2] Robi Polikar,[3] and Bahrad Sokhansanj[4]

[1] *Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA*
[2] *Spoken Language Systems Laboratory, INESC-ID, 1000 Lisbon, Portugal*
[3] *Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, USA*
[4] *School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, PA 19130, USA*

Correspondence should be addressed to Gail Rosen, gailr@ece.drexel.edu

Received 5 June 2008; Revised 19 September 2008; Accepted 30 September 2008

MetaCV

Published online 24 April 2012 Nucleic Acids Research, 2012, Vol. 40, No. 14 **e111**
doi:10.1093/nar/gks335

Rapid identification of high-confidence taxonomic assignments for metagenomic data

Norman J. MacDonald, Donovan H. Parks and Robert G. Beiko*

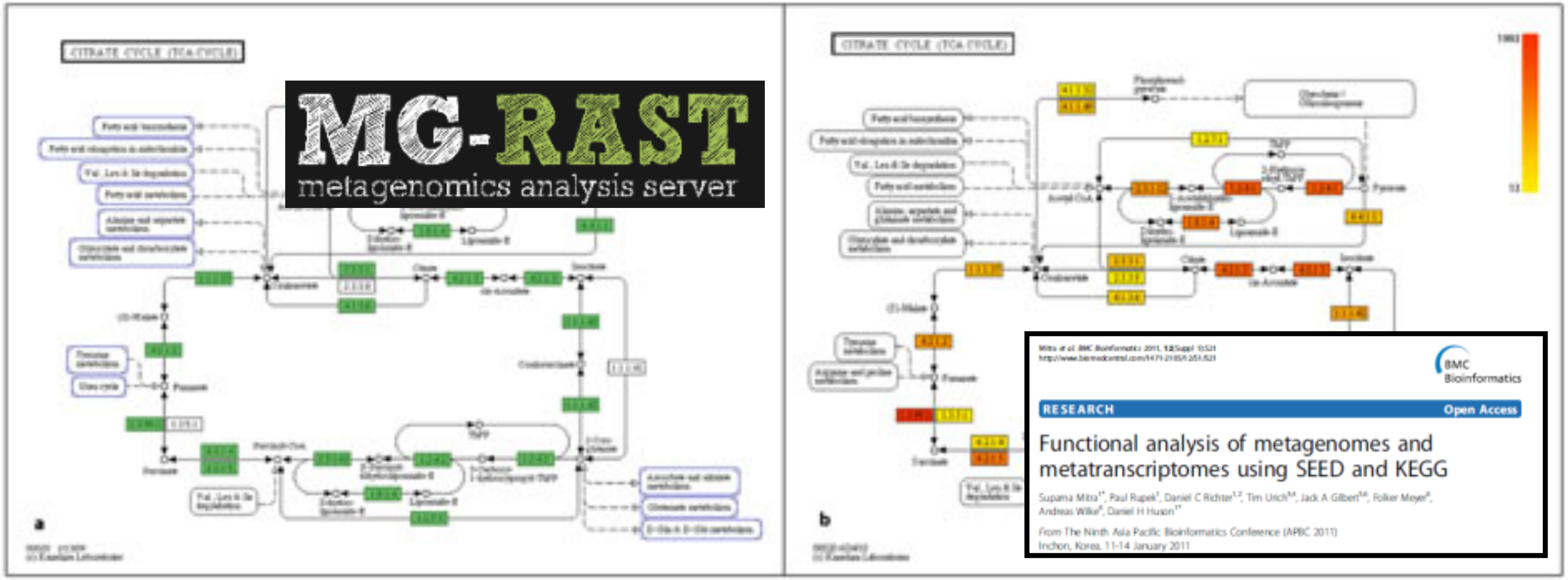
Faculty of Computer Science, Dalhousie University, 6050 University Avenue, PO BOX 15000, Halifax, NS B3H 4R2, Canada

Received December 3, 2011; Revised March 30, 2012; Accepted April 4, 2012

RITA

Visualizing results

Metabolic pathways are among the most highly conserved and best characterized systems


CITRATE CYCLE (TCA CYCLE)
MG-RAST
metagenomics analysis server
a
CITRATE CYCLE (TCA CYCLE)
b

Mitra et al. BMC Bioinformatics 2011, 12(Suppl 1):S21
http://www.biomedcentral.com/1471-2105/12/S1/S21

BMC Bioinformatics

RESEARCH Open Access

Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG

Supama Mitra[1*], Paul Rupek[1], Daniel C Richter[1,2], Tim Urich[3,4], Jack A Gilbert[5,6], Folker Meyer[6], Andreas Wilke[6], Daniel H Huson[1*]

From The Ninth Asia Pacific Bioinformatics Conference (APBC 2011)
Inchon, Korea. 11-14 January 2011

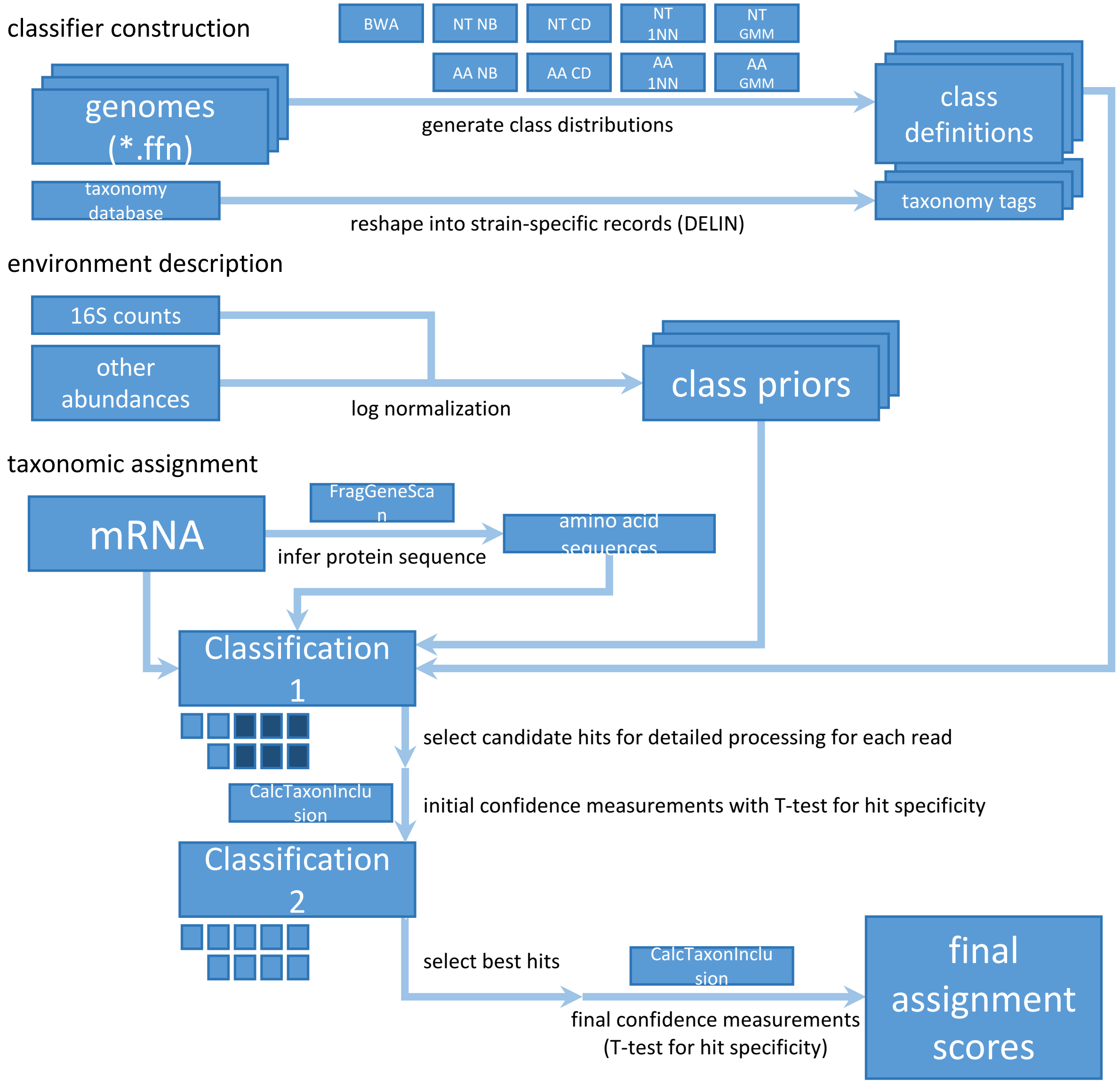
MG-RAST and MEGAN are automated metagenomic annotation tools that rely on KEGG
A major problem with KEGG pathway definitions is that the boundaries of pathways are arbitrarily defined and links between pathways (i.e. functional relationships) can be lost

Read processing – Gist

Gist is a computational pipeline for accurate assignment of reads to individual species

Integrates several methods, but uniquely assigns different weights to methods for each genome

Can also take in expected sequence distributions (e.g. based on 16S rRNA surveys)



classifier construction

BWA

NT NB

NT CD

NT 1NN

NT GMM

AA NB

AA CD

AA 1NN

AA GMM

genomes (*.ffn)

generate class distributions

class definitions

taxonomy database

reshape into strain-specific records (DELIN)

taxonomy tags

environment description

16S counts

other abundances

log normalization

class priors

taxonomic assignment

mRNA

FragGeneScan

infer protein sequence

amino acid sequences

Classification 1

select candidate hits for detailed processing for each read

CalcTaxonInclusion

initial confidence measurements with T-test for hit specificity

Classification 2

select best hits

CalcTaxonInclusion

final confidence measurements (T-test for hit specificity)

final assignment scores

Statistical considerations

There is no dedicated software or statistical tool for statistical comparisons of metatranscriptomic datasets

- Number of biological replicates? (preferably at least three)
- Differential expression of individual genes
- Gene set enrichment analyses

Ultimately metatranscriptomics could be viewed as hypothesis generating requiring subsequent targeted validation

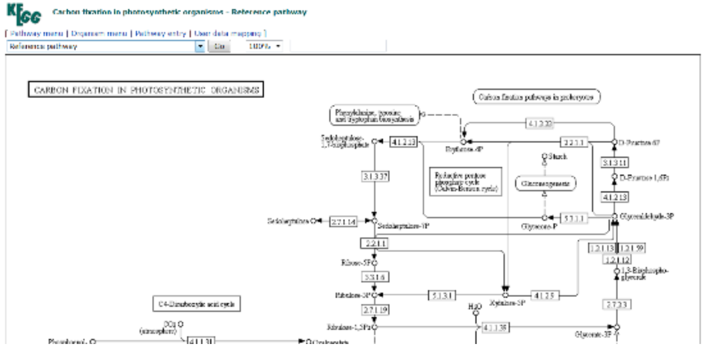
Statistical considerations

While there are no dedicated tools for metatranscriptomics analyses, tools used for RNA Seq offer potential

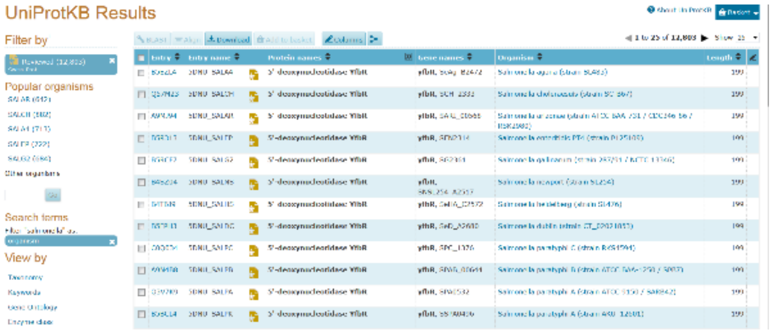
- DESeq, EdgeR, ALDEx
- Alternatively simply rely on fold change (Gfold)
- Challenges include which genes to include (minimum RPKM?)

Differentially expressed genes can be subsequently analysed through Gene Set Enrichment Approaches

Resources (Function)


KEGG
Carbon fixation in photosynthetic organisms - Reference pathway
[Pathway menu | Organism menu | Pathway entry | User data mapping]
Reference pathway
Go
100%
CARBON FIXATION IN PHOTOSYNTHETIC ORGANISMS
Carbon fixation pathways in prokaryotes
Phenylalanine, tyrosine and tryptophan biosynthesis
Reductive pentose phosphate cycle (Calvin-Benson cycle)
Starch
Gluconeogenesis
Erythrose-4P
D-Fructose-6P
D-Fructose-1,6P2
Glyceraldehyde-3P
Sedoheptulose-7P
Ribose-5P
Ribulose-5P
Xylulose-5P
Ribulose-1,5P2
Glycerate-3P
C4-Dicarboxylic acid cycle
Phosphoenol-
Oxaloacetate

KEGG


UniProtKB Results

Filter by

Reviewed (12,803)

Popular organisms

SALAR (612)
SALCH (882)
SALA4 (717)
SALEP (722)
SALG2 (684)

Other organisms

Search terms

Filter "salmonella" as organism

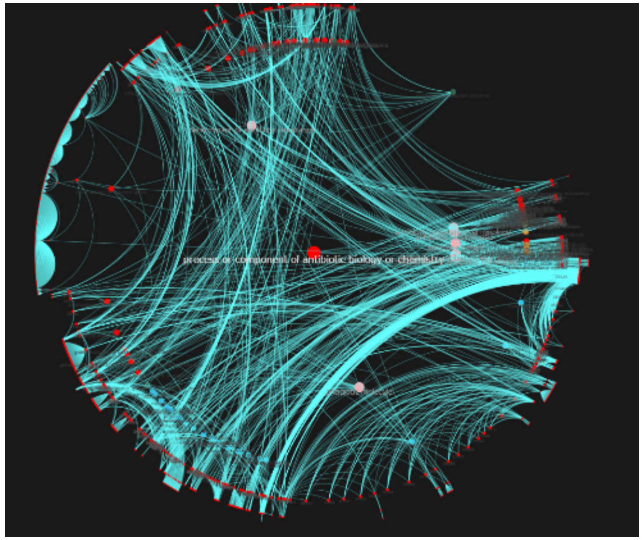
View by

Taxonomy
Keywords
Gene Ontology
Enzyme class

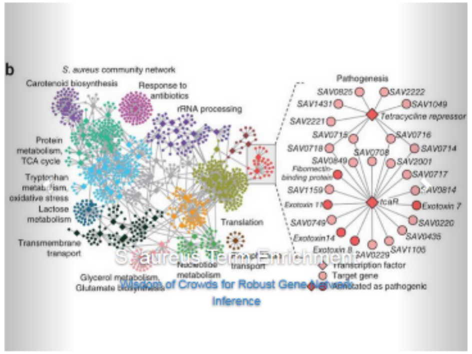
1 to 25 of 12,803 ▶ Show 25

□	Entry	Entry name	Protein names	Gene names	Organism	Length
□	B5F2L4	5DNU_SALA4	5'-deoxynucleotidase YfbR	yfbR, SeAg_B2472	Salmonella agona (strain SL483)	199
□	Q57M23	5DNU_SALCH	5'-deoxynucleotidase YfbR	yfbR, SCH_2333	Salmonella choleraesuis (strain SC-B67)	199
□	A9MJ94	5DNU_SALAR	5'-deoxynucleotidase YfbR	yfbR, SARI_00565	Salmonella arizonae (strain ATCC BAA-731 / CDC346-86 / RSK2980)	199
□	B5R3J3	5DNU_SALEP	5'-deoxynucleotidase YfbR	yfbR, SEN2314	Salmonella enteritidis PT4 (strain P125109)	199
□	B5RCF2	5DNU_SALG2	5'-deoxynucleotidase YfbR	yfbR, SG2361	Salmonella gallinarum (strain 287/91 / NCTC 13346)	199
□	B4SZ34	5DNU_SALNS	5'-deoxynucleotidase YfbR	yfbR, SNSL254_A2517	Salmonella newport (strain SL254)	199
□	B4TFN6	5DNU_SALHS	5'-deoxynucleotidase YfbR	yfbR, SeHA_C2572	Salmonella heidelberg (strain SL476)	199
□	B5FRH1	5DNU_SALDC	5'-deoxynucleotidase YfbR	yfbR, SeD_A2680	Salmonella dublin (strain CT_02021853)	199
□	C0Q5Z4	5DNU_SALPC	5'-deoxynucleotidase YfbR	yfbR, SPC_1326	Salmonella paratyphi C (strain RKS4594)	199
□	A9N4B8	5DNU_SALPB	5'-deoxynucleotidase YfbR	yfbR, SPAB_00641	Salmonella paratyphi B (strain ATCC BAA-1250 / SPB7)	199
□	Q5PIW0	5DNU_SALPA	5'-deoxynucleotidase YfbR	yfbR, SPA0532	Salmonella paratyphi A (strain ATCC 9150 / SARB42)	199
□	B5BCT4	5DNU_SALPK	5'-deoxynucleotidase YfbR	yfbR, SSPA0496	Salmonella paratyphi A (strain AKU_12601)	199

UniProtKB



CARD



Gene Ontology