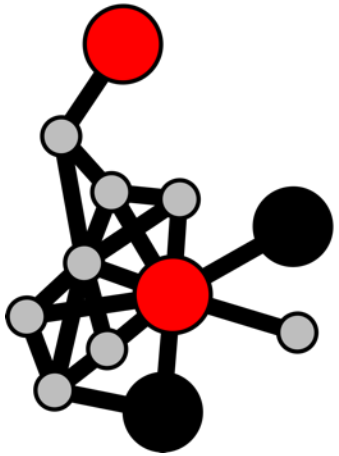
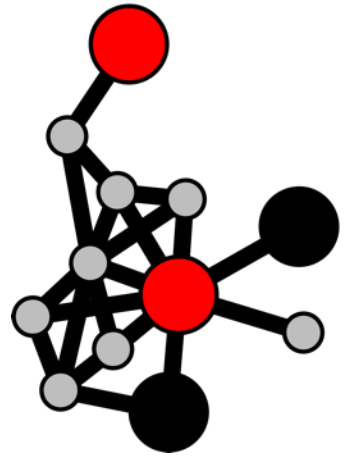
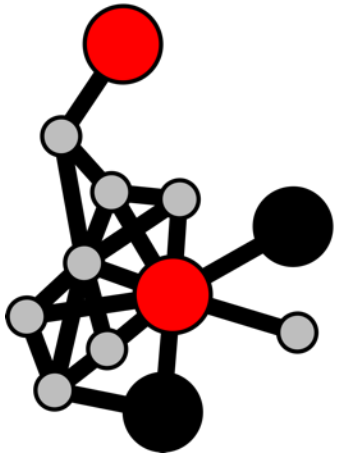
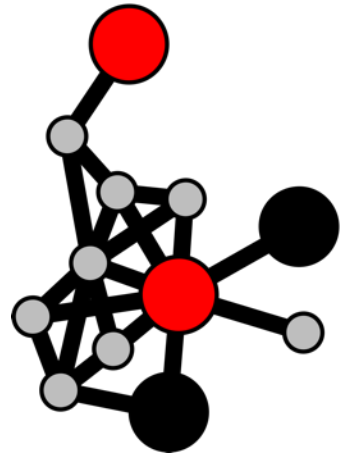
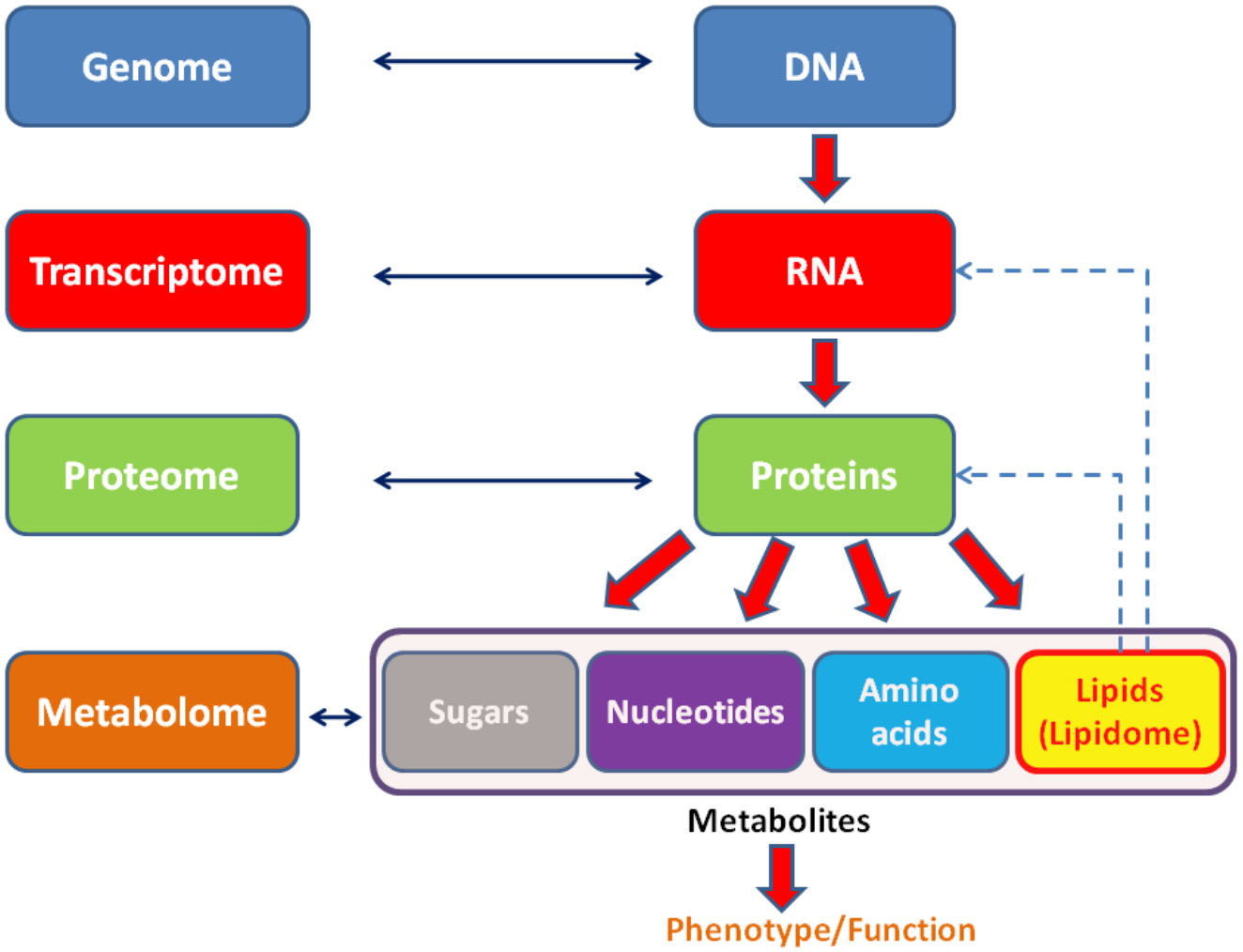





Transcriptomics lesson

Introduction

What is a transcriptome? Power and limitations



Wikipedia, User:Lmaps

The use of transcriptomics: some case studies

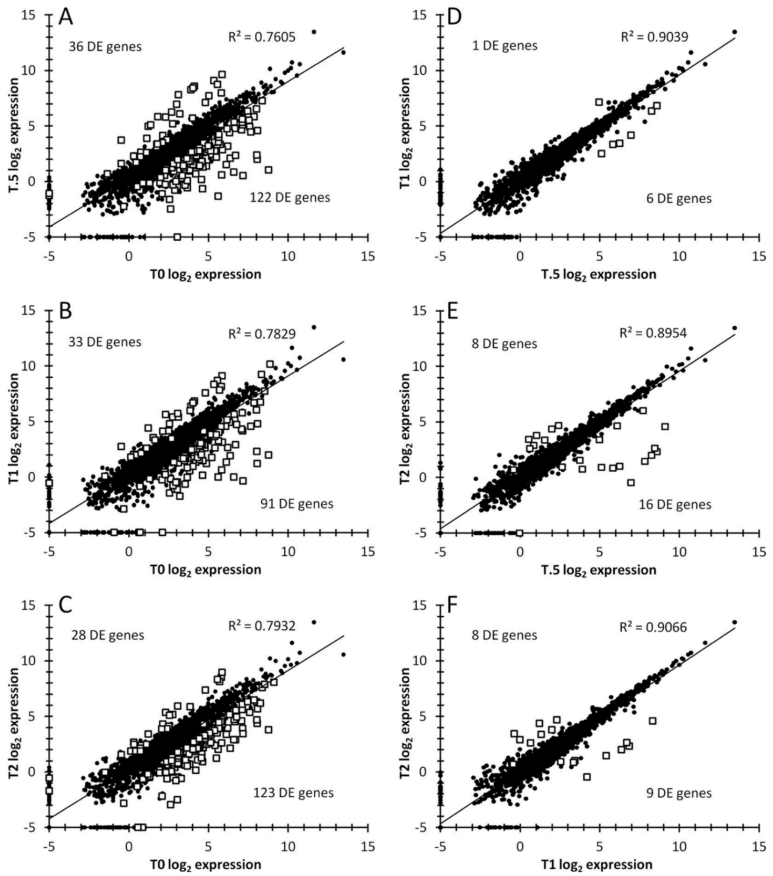
Differential gene expression analysis

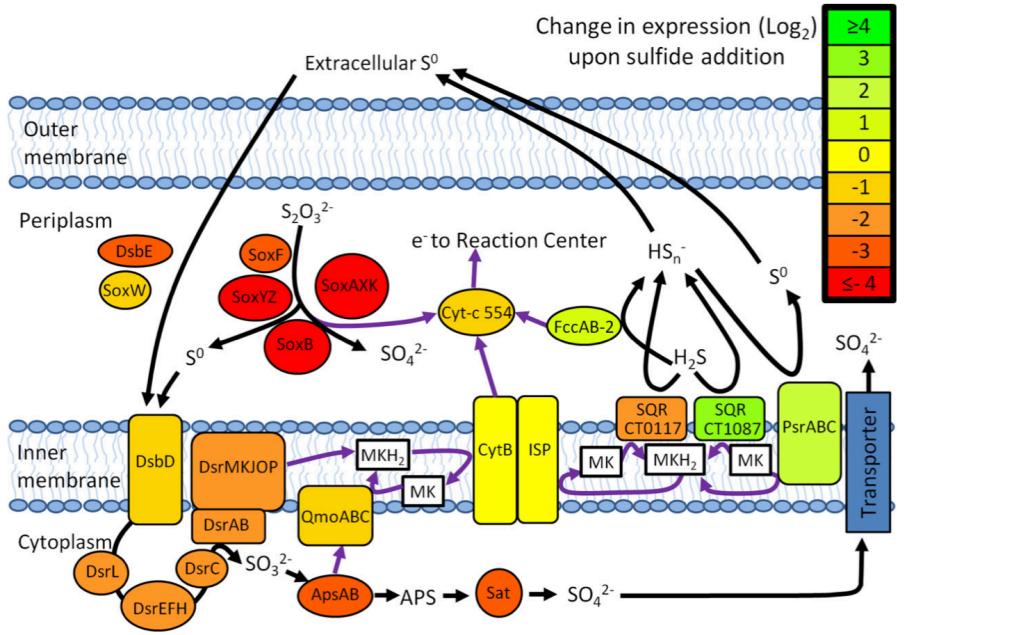
JB Journals.ASM.org

Chlorobaculum tepidum TLS Displays a Complex Transcriptional Response to Sulfide Addition

Brian J. Eddie, Thomas E. Hanson

College of Earth, Ocean, and Environment and Delaware Biotechnology Institute, University of Delaware, Newark, Delaware, USA





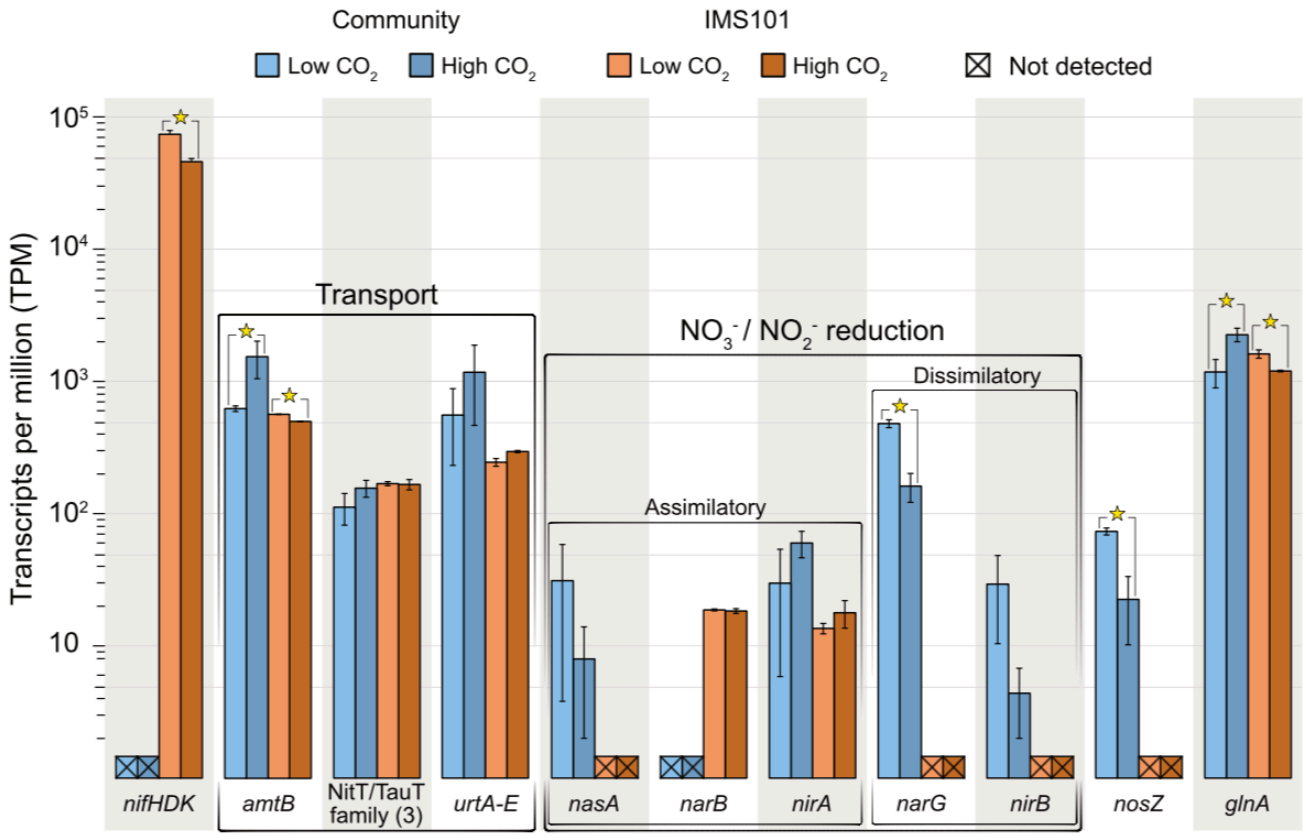
AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

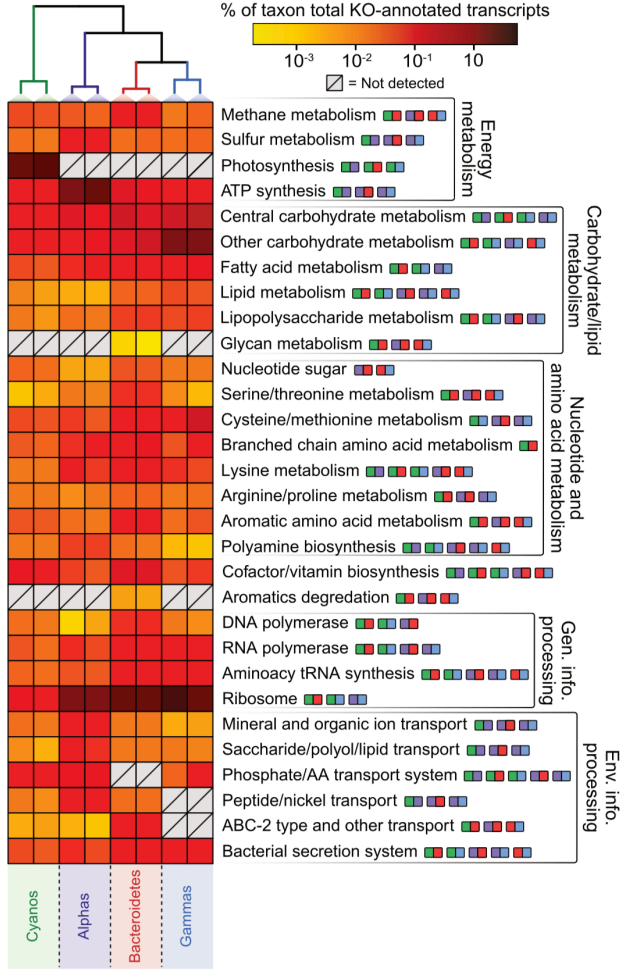
ENVIRONMENTAL MICROBIOLOGY

Gene expression by a microbial community = metatranscriptomics

Transcriptional Activities of the Microbial Consortium Living with the Marine Nitrogen-Fixing Cyanobacterium *Trichodesmium* Reveal Potential Roles in Community-Level Nitrogen Cycling

Michael D. Lee,[a] Eric A. Webb,[a] Nathan G. Walworth,[a] Fei-Xue Fu,[a] Noelle A. Held,[b] Mak A. Saito,[b] David A. Hutchins[a]





The ISME Journal (2017), 1–13

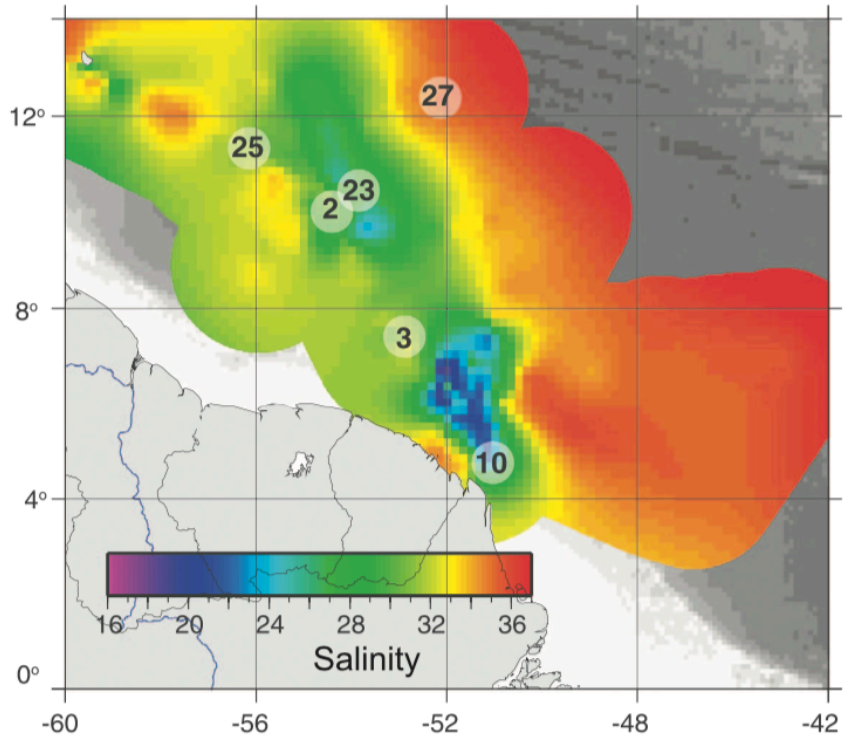
www.nature.com/ismej

ORIGINAL ARTICLE

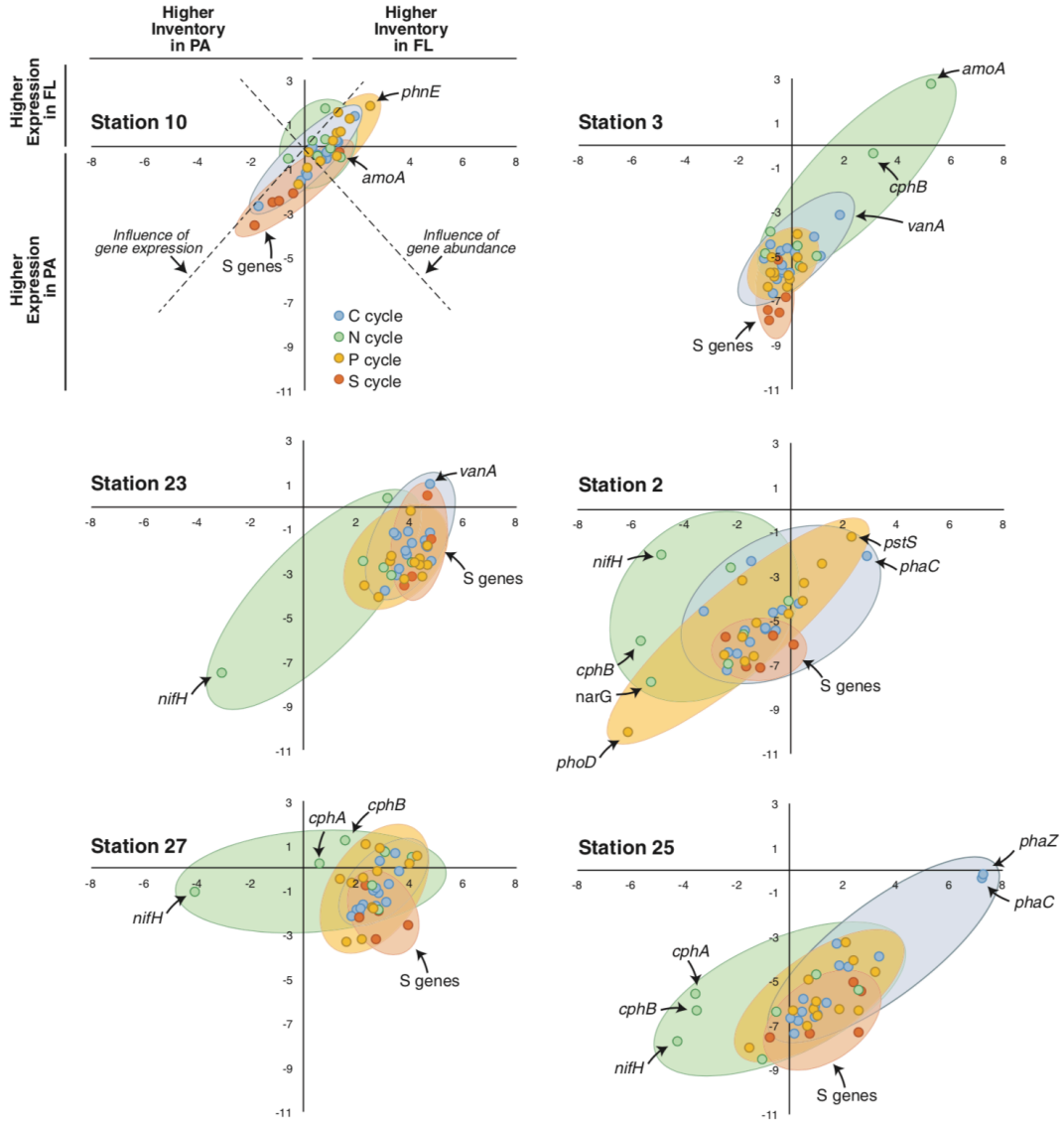
Expression patterns of elemental cycling genes in the Amazon River Plume

Brandon M Satinsky[1,5], Christa B Smith[2], Shalabh Sharma[2], Marine Landa[2], Patricia M Medeiros[2], Victoria J Coles[3], Patricia L Yager[2], Byron C Crump[4] and Mary Ann Moran[2]

[1]*Department of Microbiology, University of Georgia, Athens, GA, USA;* [2]*Department of Marine Sciences, University of Georgia, Athens, GA, USA;* [3]*Horn Point Laboratory, University of Maryland Center for Environmental Science, Cambridge, MD, USA and* [4]*College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA*



Combining metagenomics, metatranscriptomics, and differential expression analysis



The ISME Journal (2016) 10, 2106–2117

www.nature.com/ismej

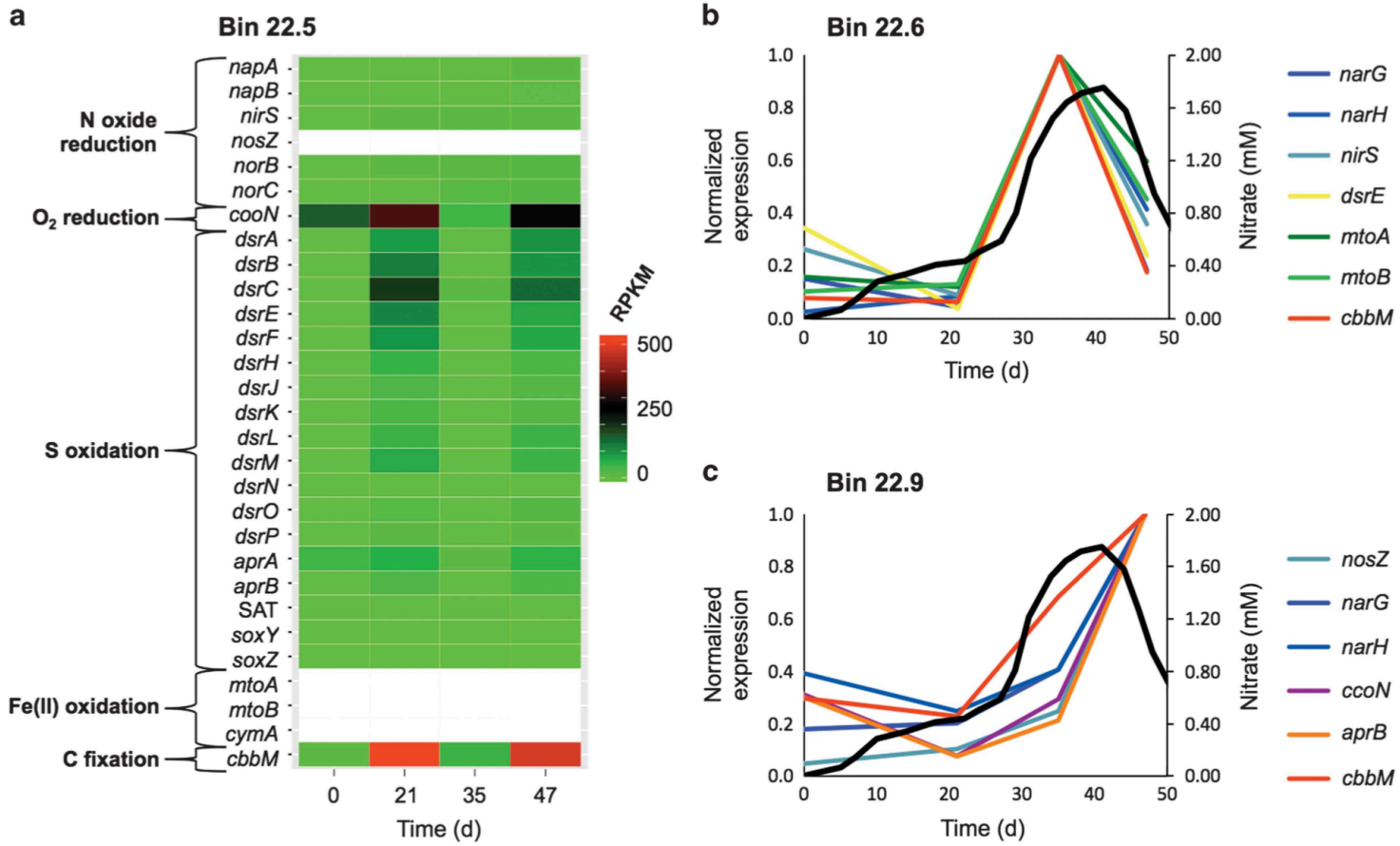
OPEN

Combining metatranscriptomics and metagenomic binning

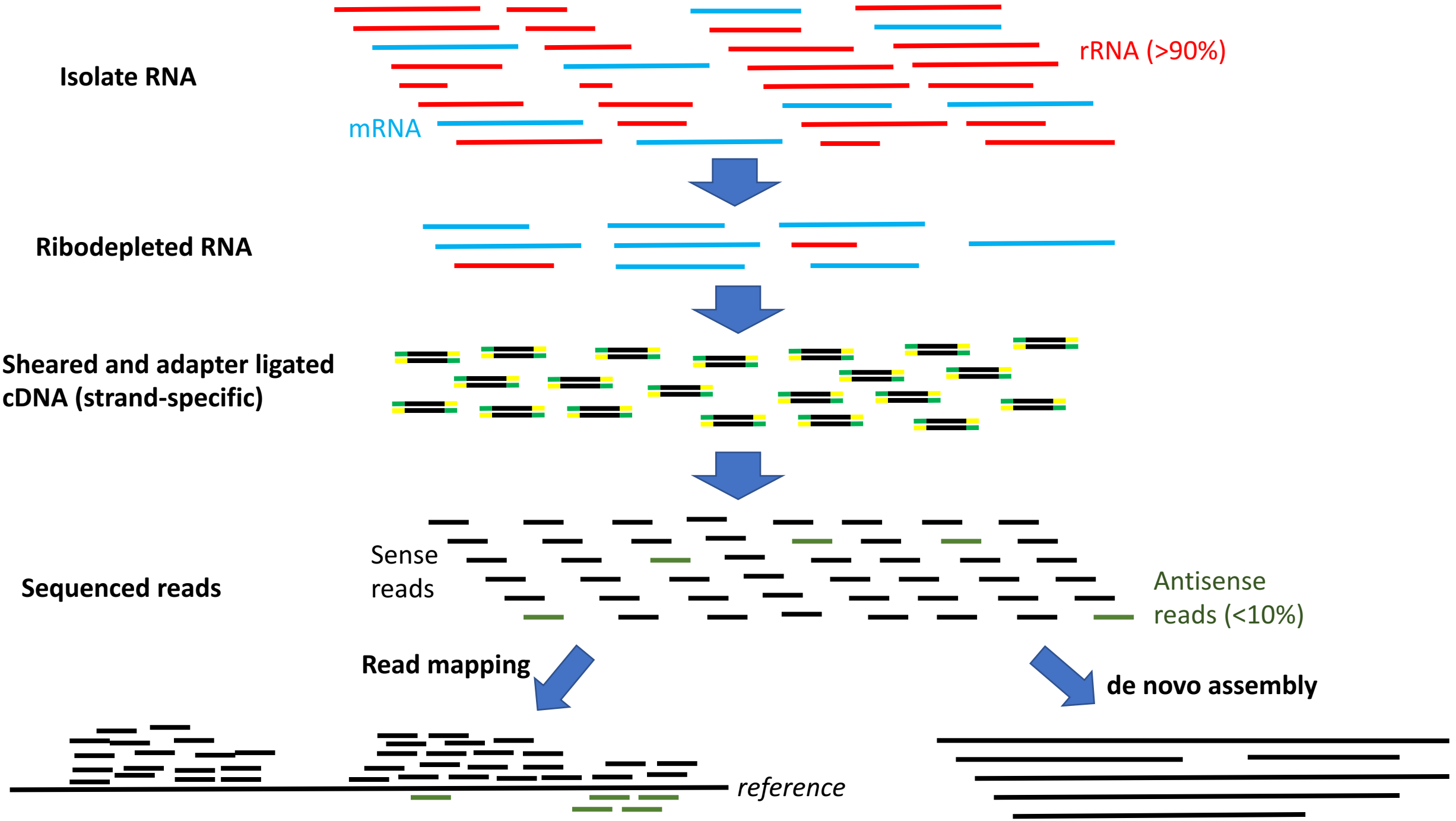
ORIGINAL ARTICLE

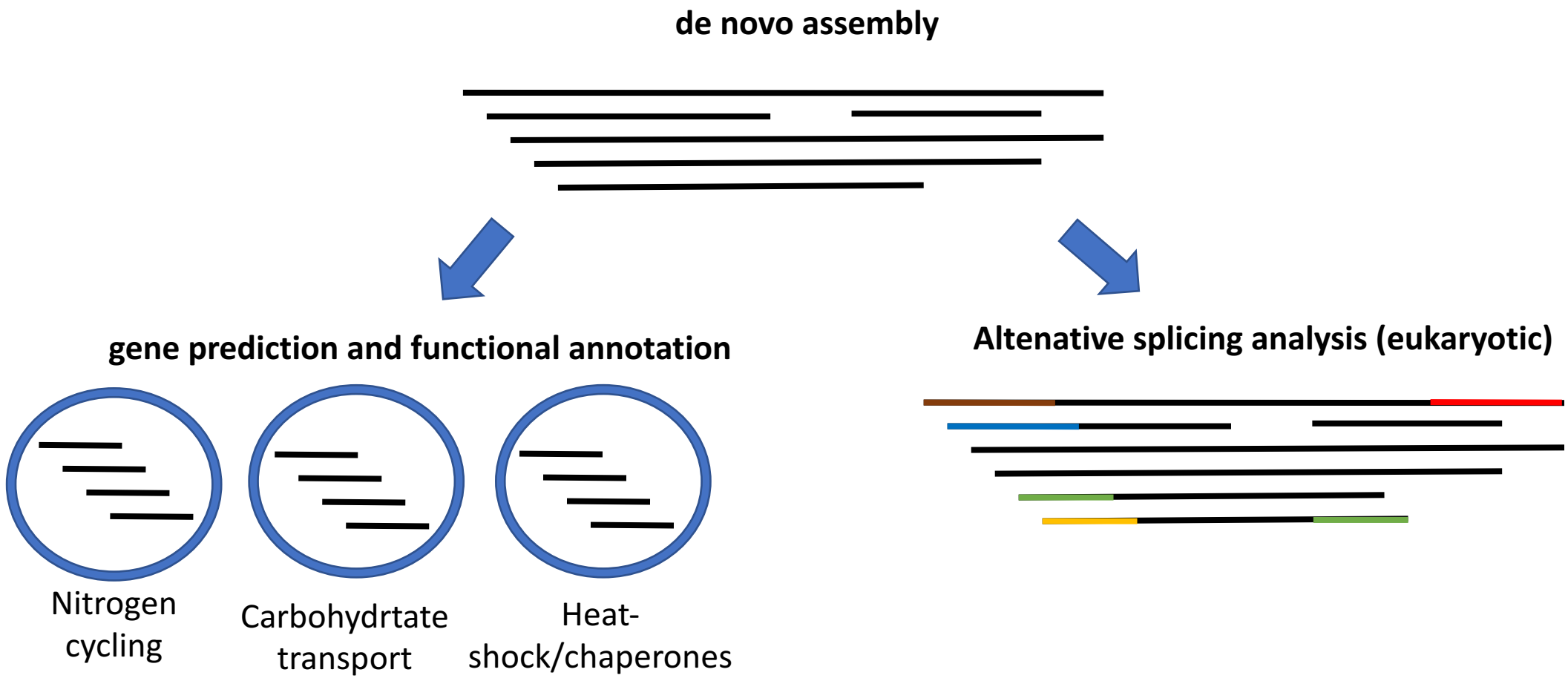
Metatranscriptomic evidence of pervasive and diverse chemolithoautotrophy relevant to C, S, N and Fe cycling in a shallow alluvial aquifer

Talia NM Jewell[1], Ulas Karaoz[1], Eoin L Brodie, Kenneth H Williams and Harry R Beller
Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA



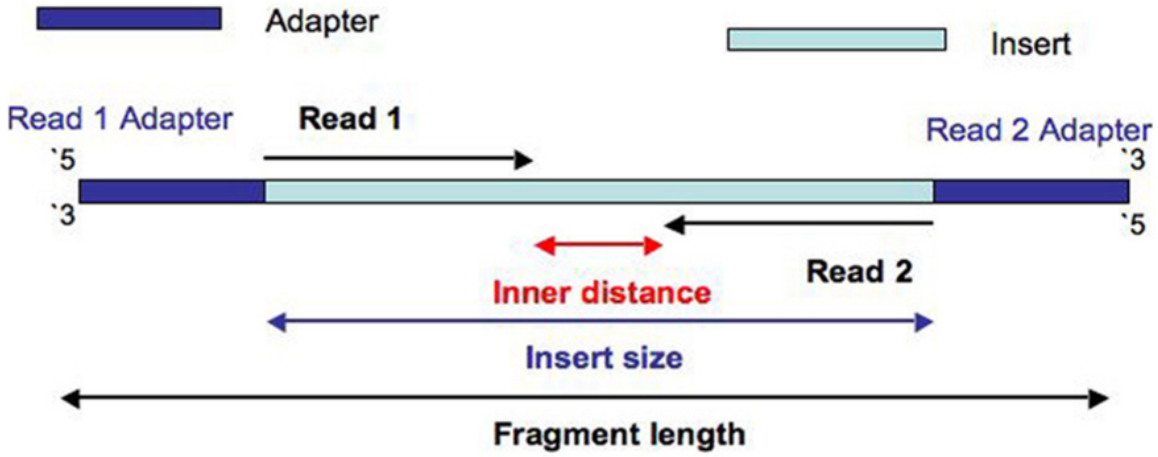
General pipeline overview


Isolate RNA
rRNA (>90%)
mRNA
Ribodepleted RNA
Sheared and adapter ligated cDNA (strand-specific)
Sequenced reads
Sense reads
Antisense reads (<10%)
Read mapping
de novo assembly
reference


de novo assembly
gene prediction and functional annotation
Altenative splicing analysis (eukaryotic)
Nitrogen cycling
Carbohydrtate transport
Heat-shock/chaperones

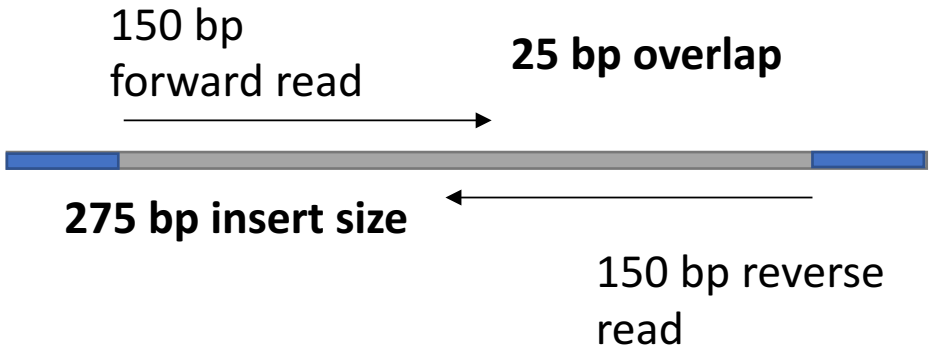
Deciding the insert size – go with overlapping read pairs

Genomics



Pierre Lindenbaum (biostars.org)

Transcriptomics



Deciding the sequencing depth

- Total reference sequence size (e.g. 3 Mb for a typical bacterial genome)
- Desired read depth for differential expression analysis (e.g. ~100x)

Averaged over the genome or housekeeping genes (if known)

- 3,000,000 x 100 = 300,000,000

Genome size

Desired coverage

- 300,000,000 / 275 = **1.1 million reads**

Insert size

Better to oversequence and then subsample. If a sample is undersequenced, you will have to seqeunce more

Transcript assemblers

Software	Released	Last updated	Computational efficiency	Strengths and weaknesses
Velvet-Oases[117][118]	2008	2011	Low, single-threaded, high RAM requirement	The original short read assembler. It is now largely superseded.
SOAPdenovo-trans[108]	2011	2014	Moderate, multi-thread, medium RAM requirement	An early example of a short read assembler. It has been updated for transcriptome assembly.
Trans-ABySS[119]	2010	2016	Moderate, multi-thread, medium RAM requirement	Suited to short reads, can handle complex transcriptomes, and an MPI-parallel version is available for computing clusters.
Trinity[120][96]	2011	2017	Moderate, multi-thread, medium RAM requirement	Suited to short reads. It can handle complex transcriptomes but is memory intensive.
miraEST[121]	1999	2016	Moderate, multi-thread, medium RAM requirement	Can process repetitive sequences, combine different sequencing formats, and a wide range of sequence platforms are accepted.
Newbler[122]	2004	2012	Low, single-thread, high RAM requirement	Specialised to accommodate the homo-polymer sequencing errors typical of Roche 454 sequencers.
CLC genomics workbench[123]	2008	2014	High, multi-thread, low RAM requirement	Has a graphical user interface, can combine diverse sequencing technologies, has no transcriptome-specific features, and a licence must be purchased before use.
SPAdes[124]	2012	2017	High, multi-thread, low RAM requirement	Used for transcriptomics experiments on single cells.
RSEM[125]	2011	2017	High, multi-thread, low RAM requirement	Can estimate frequency of alternatively spliced transcripts. User friendly.
StringTie[97][126]	2015	2019	High, multi-thread, low RAM requirement	Can use a combination of reference-guided and *de novo* assembly methods to identify transcripts.

Read alingers (there's a lot of them)

Short (unspliced) aligners [edit]

Short aligners are able to align continuous reads (not containing gaps result of splicing) to a genome of reference. Basically, there are two types: 1) based on the Burrows-Wheeler transform method such as Bowtie and BWA, and 2) based on Seed-extend methods, Needleman-Wunsch or Smith-Waterman algorithms. The first group (Bowtie and BWA) is many times faster, however some tools of the second group tend to be more sensitive, generating more correctly aligned reads.

- **BFAST** aligns short reads to reference sequences and presents particular sensitivity towards errors, SNPs, insertions and deletions. BFAST works with the Smith-Waterman algorithm. See also *seqanwers/BFAST*.
- **Bowtie** is a fast short aligner using an algorithm based on the Burrows-Wheeler transform and the FM-index. Bowtie tolerates a small number of mismatches.
- **Bowtie2** Bowtie 2 is a memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly recommended for aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.
- **Burrows-Wheeler Aligner (BWA)** BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.
- **Short Oligonucleotide Analysis Package (SOAP)**
- **GNUMAP** performs alignment using a probabilistic Needleman-Wunsch algorithm. This tool is able to handle alignment in repetitive regions of a genome without losing information. The output of the program was developed to make possible easy visualization using available software.
- **Maq** first aligns reads to reference sequences and after performs a consensus stage. On the first stage performs only ungapped alignment and tolerates up to 3 mismatches.
- **Mosaik** Mosaik is able to align reads containing short gaps using Smith-Waterman algorithm, ideal to overcome SNPs, insertions and deletions.
- **NovoAlign (commercial)** is a short aligner to the Illumina platform based on Needleman-Wunsch algorithm. It is able to deal with bisulphite data. Output in SAM format.
- **PerM** is a software package which was designed to perform highly efficient genome scale alignments for hundreds of millions of short reads produced by the ABI SOLiD and Illumina sequencing platforms. PerM is capable of providing full sensitivity for alignments within 4 mismatches for 50bp SOLID reads and 9 mismatches for 100bp Illumina reads.
- **RazerS**
- **SEAL** uses a MapReduce model to produce distributed computing on clusters of computers. Seal uses BWA to perform alignment and Picard MarkDuplicates to detection and duplicate read removal.
- **segemehl**
- **SeqMap**
- **SHRiMP** employs two techniques to align short reads. Firstly, the q-gram filtering technique based on multiple seeds identifies candidate regions. Secondly, these regions are investigated in detail using Smith-Waterman algorithm.
- **SMALT**
- **Stampy** combines the sensitivity of hash tables and the speed of BWA. Stampy is prepared to alignment of reads containing sequence variation like insertions and deletions. It is able to deal with reads up to 4500 bases and presents the output in SAM format.
- **Subread** [41] is a read aligner. It uses the seed-and-vote mapping paradigm to determine the mapping location of the read by using its largest mappable region. It automatically decides whether the read should be globally mapped or locally mapped. For RNA-seq data, Subread should be used for the purpose of expression analysis. Subread can also be used to map DNA-seq reads.
- **ZOOM (commercial)** is a short aligner of the Illumina/Solexa 1G platform. ZOOM uses extended spaced seeds methodology building hash tables for the reads, and tolerates mismatches and insertions and deletions.
- **WHAM** WHAM is a high-throughput sequence alignment tool developed at University of Wisconsin-Madison. It aligns short DNA sequences (reads) to the whole human genome at a rate of over 1500 million 60bit/s reads per hour, which is one to two orders of magnitudes faster than the leading state-of-the-art techniques.

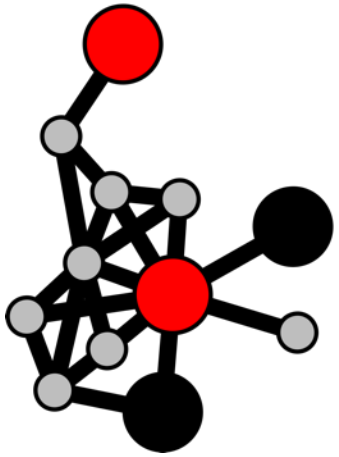
Spliced aligners [edit]

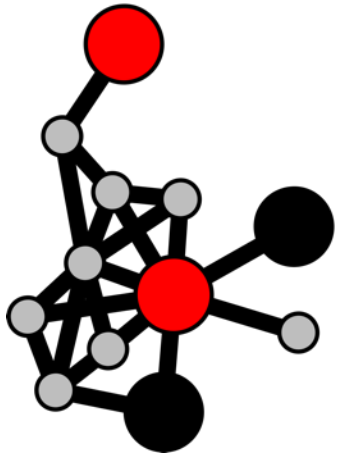
Many reads span exon-exon junctions and can not be aligned directly by Short aligners, thus specific aligners were necessary - Spliced aligners. Some Spliced aligners employ Short aligners to align firstly unspliced/continuous reads (exon-first approach), and after follow a different strategy to align the rest containing spliced regions - normally the reads are split into smaller segments and mapped independently. See also.[42][43]

Aligners based on known splice junctions (annotation-guided aligners) [edit]

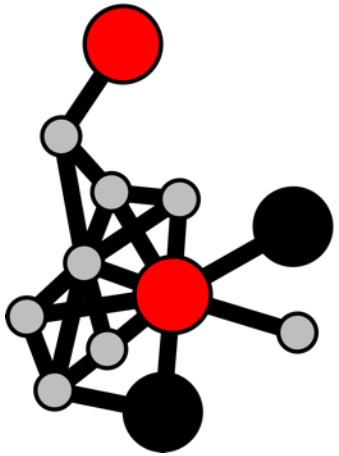
In this case the detection of splice junctions is based on data available in databases about known junctions. This type of tools cannot identify new splice junctions. Some of this data comes from other expression methods like expressed sequence tags (EST).

- **Erange** is a tool to alignment and data quantification to mammalian transcriptomes.
- **IsoformEx**
- **MapAL**
- **OSA**
- **RNA-MATE** is a computational pipeline for alignment of data from Applied Biosystems SOLID system. Provides the possibility of quality control and trimming of reads. The genome alignments are performed using **mapreads** and the splice junctions are identified based on a library of known exon-junction sequences. This tool allows visualization of alignments and tag counting.
- **RUM** performs alignment based on a pipeline, being able to manipulate reads with splice junctions, using Bowtie and Blat. The flowchart starts doing alignment against a genome and a transcriptome database executed by Bowtie. The next step is to perform alignment of unmapped sequences to the genome of reference using BLAT. In the final step all alignments are merged to get the final alignment. The input files can be in FASTA or FASTQ format. The output is presented in RUM and SAM format.
- **RNASEQR**.





Some upcoming tutorials (not in that particular order)



FastQC

Trimmomatic

Flash

rRNA depletion (*in silico*)

Transcript assembly

Read mapping and estimation of expression