Computational Molecular Biology (BIOMEDIN231) Final Project
Yooli Light (SCPD, Stanford ID#05929351)

Metatranscriptomic Pipeline Draft for Linking Functions to Microbes in Environmental Microbial Community

Abstract

In the advances of sequencing and computing technologies, metatranscriptome RNA-seq approach to understanding microbial community and functions in terms of biological processes is possible. However, the resolution of the metatranscriptomic analysis in microbial community is limited to microbe composition and functional or metabolic potential of the community biased in existing reference genomes of microbiome. Higher resolution of relationship between functions and individual microbe's functional blueprints would provide better understanding of interaction of the community with their environments. Here I propose a draft of metatranscriptomic pipeline that possibly connect microbes to functions in the community and provide a strategy to improve metatranscriptomic analysis by incorporating single cell bead pipeline to a metatransciptomic pipeline utilizing exiting tools and pipelines for metagenomics and metatranscriptomics.

Overview the metatranscriptomic pipeline

The overview of the pipeline is described in Figure 1. It consists of four components: phylogenic profiling of the community, metagenome analysis, metatranscriptome analysis and single cell analysis. A detailed description of each component follows.

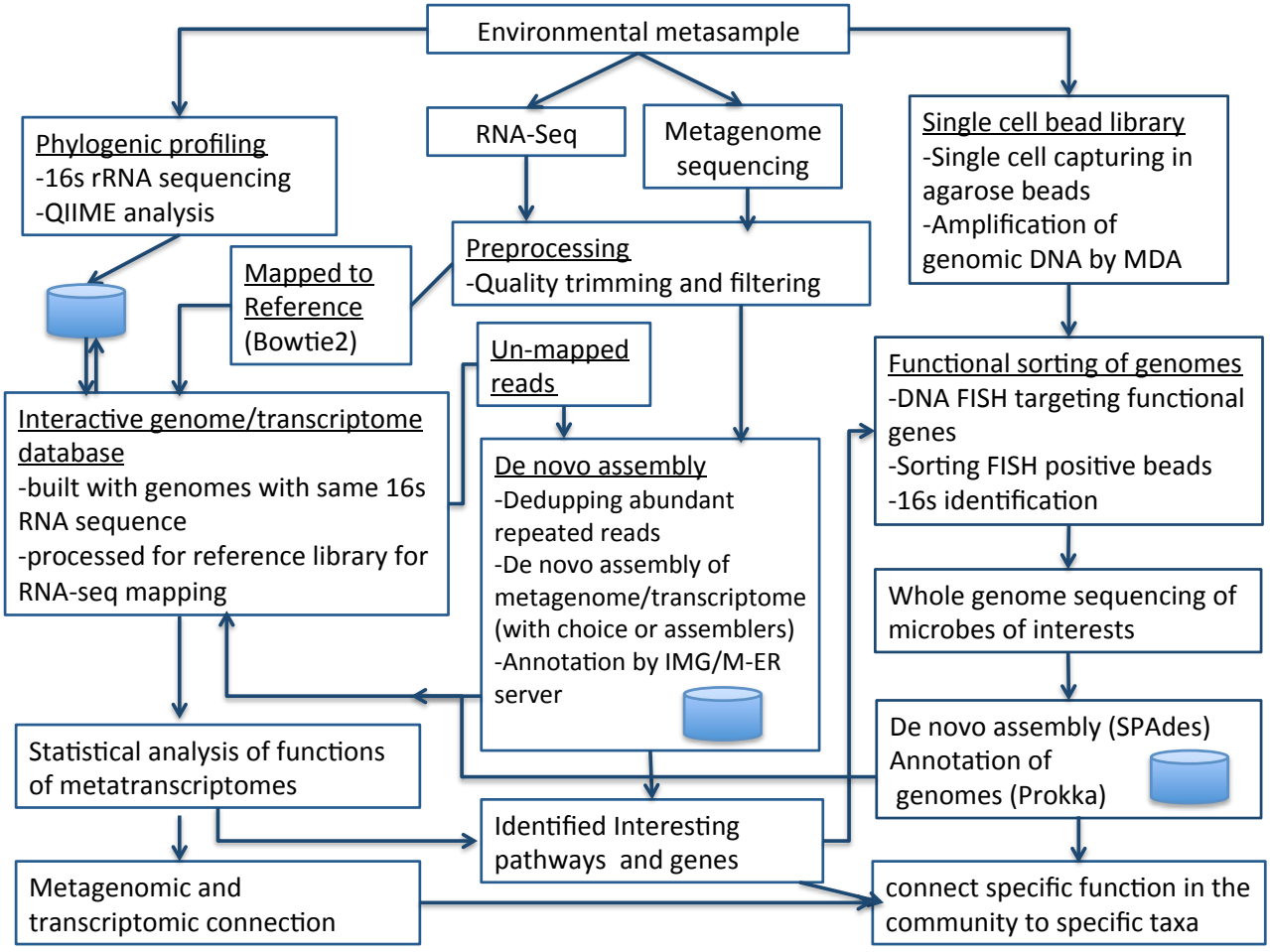

Figure1. Summary of Metatranscriptomic Pipeline for Linking Functions to Microbes in Environmental Microbial Community. It consists of four components: phylogenic profiling of the community, metagenome analysis, metatranscriptome analysis and single cell analysis. Phylogenic profiling by 16s rRNA DNA is performed prior to the rest of three analyses and interactive database containing reference genomes and transcriptomes are built based on the 16s profile analysis. Metagenome sequencing followed by assembly and annotation is performed and annotated metagenome contigs/scaffolds are added to the database. The sequencing reads from RNA-seq data sets are first mapped to the reference sequences in the database and un-mapped sequences are fed into de novo assembly/annotation pipeline. The annotated transcript contigs are deposited as reference sequences. Statistical analysis of expression by counting mapped reads to references identifies genes of interesting functions or in the pathways. The genomes containing the genes are labeled by FISH using the gene probes and sorted for sequencing. The single amplified genomes, which are sorted by functions of interest, are assembled with single cell genome assembler, annotated, and also deposited into the database.

16 s rRNA profiling to identify microbes in the community

Prior to functional analysis of community, phylogenic profiling of microbial communities of interest is performed to estimate the complexity and diversity of the community and to compare among the communities. The implemented method was chosen based on the requirements that include short turn-out time from sequencing to data analysis (within 2days), multiflexing capability and built-in statistical analysis tools to be able to compare communities. The method for ultra-high-throughput microbial community analysis using Illumina MiSeq platform (Caporaso, Lauber et al. 2012) was chosen for this pipeline. Briefly, V4 region of 16S rRNA DNA sequences is amplified with Illumina adapters and barcodes and the paired-end reads are quality filtered and processed through QIIME, an acronym for Quantitative Insights Into Microbial Ecology (Caporaso, Kuczynski et al. 2010). MiSeq generates around 5 million 150 bp paired-end reads (2x 150bp) in a day, which is suitable for primary investigation for metagenomic samples. QIIME is an open-source bioinformatics software package designed for microbial community analysis based on DNA sequence data and assigns operational taxanomic units accurately and provides phylogenic and taxon-based analysis of diversity within and between samples with nice visualization (Caporaso, Kuczynski et al. 2010). The manual and detail information is well documented in *Curr Protoc Microbiol* Chapter 1 (Kuczynski, Stombaugh et al. 2012).

Interactive database

Also, these profiling data will be served as queries to bring genomes for building interactive genome database. The idea is that microbial genomes and metagenomes that are identified in the 16 profiling results are brought to the database and processed as reference genome database for RNA-seq alignment, contributing to speed and efficiency. In addition, custom databases generated with data sets from the sample used for metatranscriptomics can be used for more accurate statistical analysis of expression by relative frequencies of transcript comparison to the

database. The assembled and annotated metagenomes, RNA-seq contigs and single cell genomes are deposited into this database as reference sequences, which is served as references for the iterative mapping of the short sequencing reads.

RNA-seq and metagenonme process pipelines

The RNA-seq process pipeline consists of three major components: preprocessing, assembly, and annotation.

Preprocessing: Quality score based trimming and filtering

Preprocessing steps including quality trimming and filtering and ribosomal RNA subtraction improves the accuracy and computational efficiency of both mapping and *de novo* assembly of metatranscriptomes. Artifacts from RNA-seq data such as sequencing adaptors and low-complexity reads can be removed using tools, such as TagDust (Lassmann T Fau - Hayashizaki, Hayashizaki Y Fau - Daub et al.) and SeqTrim (Falgueras J Fau - Lara, Lara Aj Fau - Fernandez-Pozo et al.). Sequencing errors in reads can be removed or corrected by analyzing the quality score. Ribosomal RNA (rRNA) also need to removed since it comprises over 80% of total RNA from prokaryote transciptome and regardless of rRNA removal steps in sample preparation, considerable numbers of reads are derived from rRNA (He, Wurtzel et al. 2010, Stewart, Ottesen et al. 2010).

RNA-seq Analysis with reference genomes or transcriptomes

RNA-seq analysis is based on mapping the sequencing reads to known reference genomes, annotated transcriptomes or transcripts assembled from the sequencing reads. For read mapping tools to reference sequences, two main classes of tools are available, unsliced aligner and spliced-junction aligner. The splice junction aligners are more useful for mapping eukaryotic transcriptomics and only unspliced

alignment tools were considered for the metatranscriptomic pipeline. Two main approaches are taken to align short reads to reference sequences. The first approach is hash-based methods, hashing either reads (e.g., SOAP (Li, Li et al. 2008), ELAND [Cox 2007]) or the reference genome (Novoalign [www.novocraft.com], Mosaik (Smith, Quinlan et al. 2008)). The other approach is based on Burrows-Wheeler Transform such as BWA (Li and Durbin 2009), SOAP2 (Li, Yu et al. 2009), and Bowtie (Langmead et al. 2009). The hash-based mappers can achieve better sensitivity but take more time and computational cost than Burrows-Wheeler Transform based aligners which perform very efficiently in ungapped alignments of short reads. SOAP2 and Bowtie don't allowed gaps in paired-end reads. There is also the hybrid approach such as STAMPY which improves speed without compromising sensitivity to gapped alignments (Lunter and Goodson 2011). Bowtie2 also combined the strengths of the two approaches by dividing the algorithms broadly into two stages: an ungapped seed-finding stage that benefits from the speed and memory efficiency of the full-text minute index and a gapped extension stage that uses dynamic programming (Langmead and Salzberg 2012). Bowtie2 achieves a combination of speed, sensitivity and accuracy across read lengths and sequencing technology, thus is implemented as the align tool for the pipeline.

RNA-seq Analysis without reference genomes or transcriptomes

Most of time in metagenomic and metatranscriptomic analysis, *de novo* assembly of the short reads into longer 'useful' contigs, from which information can be extracted in terms of biological processes, owing to largely un-sequenced microbial genomes. After the analysis with reference genomes, the mapped reads are filtered and the un-mapped reads are assembled into contigs using a choice of *de novo* transcriptomic assembler and annotated through a server (described in annotation section) and these annotated contigs become reference sequences that the original reads are mapped back to.

1. *De novo* assembly

De novo transcriptome assembly faces more challenges due to the sequencing coverage among different transcripts which can range over many orders of magnitude, depending on transcript abundance and sequencing depth (Martin, Bruno et al. 2010). Several algorithms are reported for genome-independent transcriptome reconstruction such as Trinity, SOAPdenovo, Trans-Abyss, and Oases, which are mostly modified from genome assembly tools except Trinity. These *de novo* assemblers uses same approach, the *de Bruijn* graph based method to reconstruct transcripts and then post-process the assembly to merge contigs and remove redundancy. The overview of this assembly strategy is well illustrated in Figure2, copied from a review paper (Martin and Wang 2011).

An optimal *de novo* assembler needs to be chosen for the generated data set with consideration of variables which can affect the performance of assembly. Zhao and colleagues compared performance of these transcriptome assemblers under important variables affecting *de novo* assembly (Zhao, Wang et al. 2011). The variables are: single k-mer vs. multiple k-mer values, single vs. multiple genome, low vs. high sequence coverage depth, non-directional vs directional reads. Also, the run time and memory required for each of the tools were compared in this study. Based on the guidelines from this comparative study, the flexibility of assembler options will be implemented so that the optimal assembler will be used depending on the characteristics of the metatranscriptomic data. Briefly, a multiple k-mer approach should be considered to achieve better assembly results. Trinity seems the best single k-mer assembler for transcriptome assembly for both small and large data sets across various conditions. But it requires a very long running time (several hours vs, hundreds of hours). Oases with multiple k-mer and trans-Abyss produces the more diverse long transcripts but Oases requires large memory (20-40G vs. 140G). SOAPdenovo uses the smallest memory and shortest runtime but in general short and incomplete transcriptomes are generated especially with large amounts of

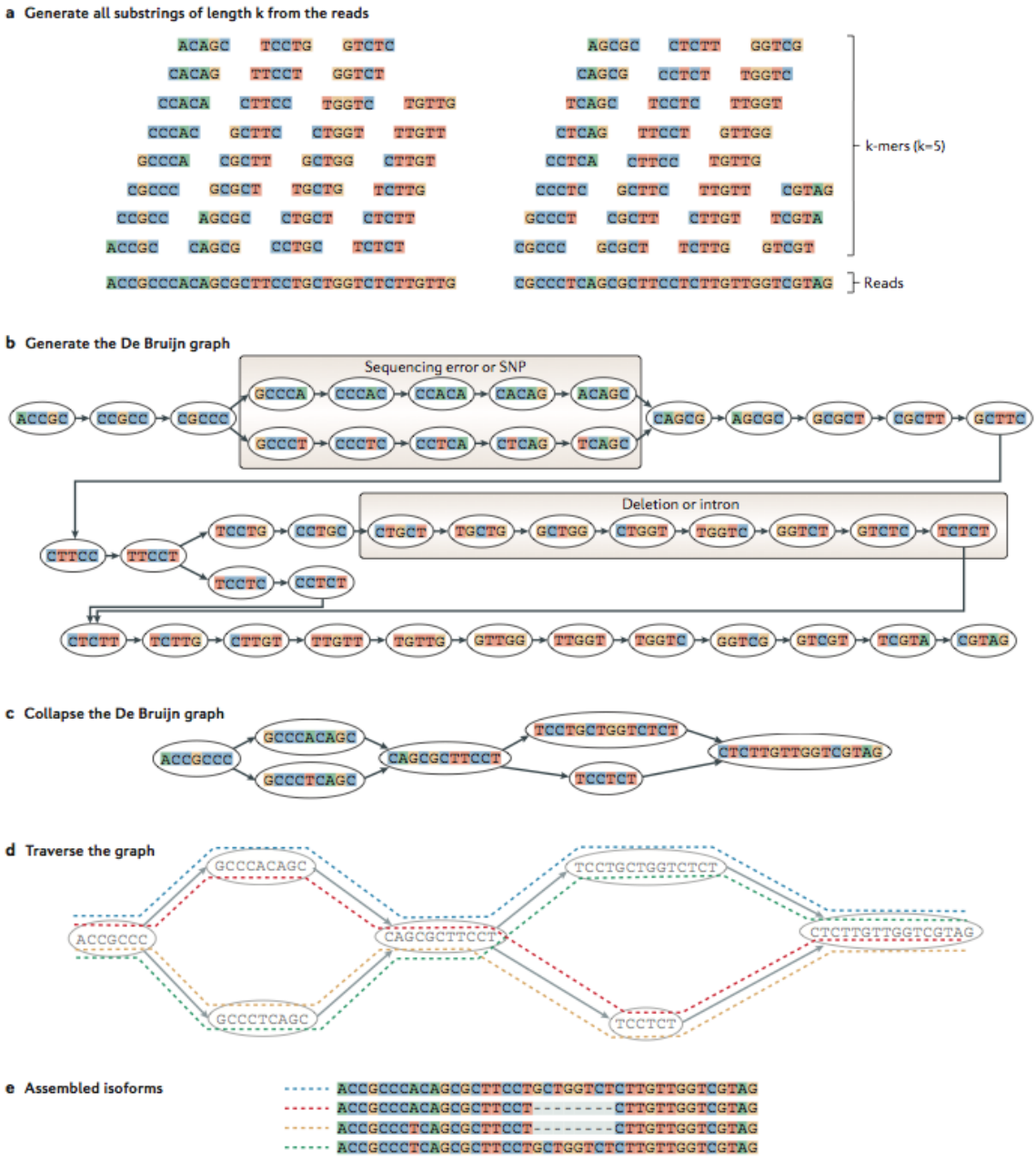

Figure2. **Overview of the *de novo* transcriptome assembly strategy. a** | All substrings of length k (k-mers) are generated from each read. **b** | Each unique k-mer is used to represent a node (or vertex) in the De Bruijn graph, and pairs of nodes are connected if shifting a k-mer by one character creates an exact k–1 overlap between the two k-mers. Note that for non-strand-specific RNA sequencing data sets, the reverse complement of each k-mer will also be represented in the graph. Here, a simple example using 5-mers is shown. The example illustrates a SNP or sequencing error (for example, A/T) and an example of an intron or a deletion. Single-nucleotide differences cause 'bubbles' of length k in the *De Brujin* graph, whereas introns or deletions introduce a shorter path in the graph. **c,d** | Chains of adjacent nodes in the graph are collapsed into a single node when the first node has an out degree of one and the second node has an in degree of one. Last, as in the reference-based approach, four alternative paths (blue, red, yellow and green) through the graph are chosen. **e** | The isoforms are then assembled.

sequence inputs with high coverage depth. Large data set can be divided into a series of 05, 1, 3G subsets to test for the optimal conditions for assembly. Usually 100x average coverage on estimated size of expressed transcripts is recommended to start with for de novo assembly (Zhao, Wang et al. 2011).

2. Assembled genome and transcriptome annotation

The assembled transcripts are assigned to the annotation pipeline. The annotation pipelines consist of mainly two subunits, gene prediction modules and functional annotation modules of predicted genes. The microbial genome annotation services are available through web-based annotation servers, NCBI PGAAP, RAST, JGI IMG and JCVI and metagenome annotation through MG-RAST and IMG/M (Markowitz, Chen et al. 2012). Functional annotation of metagenome and metatranscriptome requires a large-scale database of metagenomic data sets and substantial computational capabilities. Thus, the annotation of metagenome and metatranscriptome is performed in one of these servers. These servers have implemented a series of the existing algorithms in their pipelines and here is brief description of the tools and approaches that are used in JGI/IMG M pipelines as an example.

a. Gene prediction

The assembled transcripts are first screened by algorithms that predict non-coding RNA genes, rRNA, tRNA, other non-coding RNA and CRISPR. For tRNA prediction, ARAGORN (Laslett and Canback 2004) and tRNAscan-SE (Lowe and Eddy 1997) are the main tools of choice. For ribosomal RNA prediction, RNAmmer (Lagesen, Hallin et al. 2007), which uses HMM library created from structural alignments, reliably finds ribosomal RNA. Other non-coding gene prediction is performed using the INFRNAL software package (Nawrocki, Kolbe et al. 2009) using PFAM database of ncRNA families represented by profile SCFGs (stochastic context-free grammars) of the secondary structure and primary sequence profile and multiple sequence alignments of ncRNAs (Griffiths-Jones, Moxon et al. 2005). CRISPR identification is

performed mainly with CRT (CRISPR Recognition Tool), which can cope with unique spacer sequences separating the repeats better than existing repeat detection tools (Bland, Ramsey et al. 2007).

Prediction of protein-coding genes can be carried by two different categories, evidence-based prediction and *ab initio* CDS prediction. Evidence-based prediction searches for homologs to CDSs of known proteins. This method can predict genes more accurately than *ab initio* prediction but can miss unique genes and be sensitive to database errors. The *ab initio* CDS prediction, which can perform much faster than evidence-based prediction, takes several different approaches: heuristics and rule-based tools such as Prodigal (PROkaryotic DYnamic programming Gene-finding Algorithm) (Hyatt, Chen et al. 2010), HMM-based algorithms such as GeneMarkS (Besemer, Lomsadze et al. 2001) and GLIMMER (Salzberg, Delcher et al. 1998).

Gene prediction in metagenomics has more challenges due to the large proportion of fragmented and incomplete contigs assembled from short reads. Here are lists of algorithms that were developed for gene prediction in metagenomic projects. Orphelia (Hoff, Lingner et al. 2009) utilizes prediction models that were created with machine learning on the basis of a wide range of annotated genomes. MetaGene (Noguchi, Park et al. 2006) utilizes di-codon frequencies estimated by the GC content of a giving sequence with other various measures. FragGeneScan (Rho, Tang et al. 2010) combines sequencing error model and codon usages in a HMM to improve the prediction of protein-coding region in short reads. Gene prediction in metagenome and transcriptome data is performed in parallel with MetaGene, FragGeneScan, GeneMark and Prodigal, and then consolidated by resolving the overlaps.

b. **Functional annotation of predicted coding genes**

Predicted coding genes are assigned to product names through comparison to publically available protein family and protein databases. The basic flow of the process was adopted from the IMG gene annotation pipeline (Figure 3).

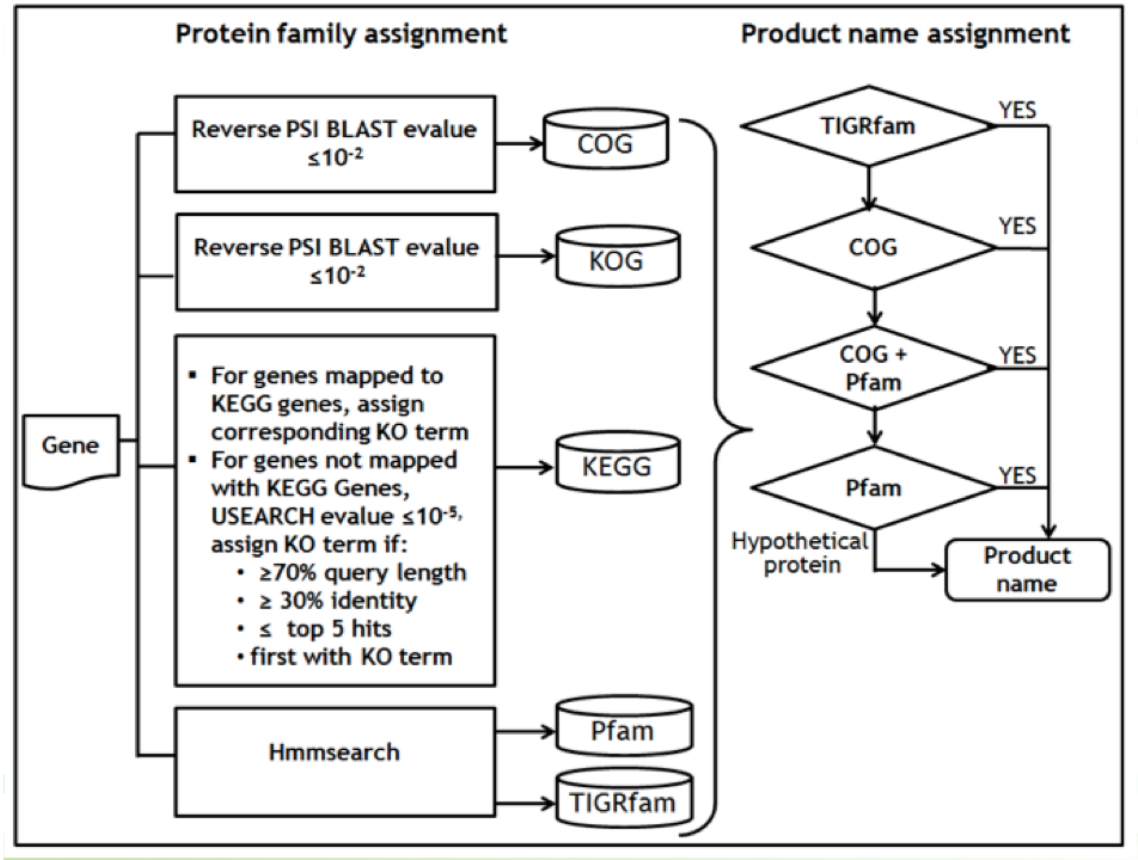

Figure 3. The gene product name assignment procedure used in the IMG pipeline. Genes are first compared to protein families and protein databases. A product name is assigned through a series of checks to identify significant hits to IMG terms and the protein families databases. At the end of the process translation tables are used to produce a Genbank compliant product name for the respective source (Standarads in Genomic Scienc, 2009).

The protein sequences of predicted genes are compared to COG PSSMS obtained from the CDD, Conserved Domain Database, (Marchler-Bauer, Anderson et al. 2007) using RPS-BLAST and only the top hit is retained with set e-value cutoff. Also, the sequences are searched against the KEGG gene database (Kanehisa, Araki et al. 2008) using BLASTp and an e-value cutoff of 1e-5. A KEGG Orthology rank of 5 or

better is assigned and greater than 70% alignment length on the query and KEGG gene sequences with the top hit recorded. Then, the sequence are searched against the Pfam (Punta, Coggill et al. 2012) and TIGfam database (Haft, Selengut et al. 2003) using a BLAST prefiltering and subsequent comparison to HMMs using hmmsearch (Eddy 1998).

Functional annotation in metagenome/metatranscriptome relies on classifying sequences to known functions or taxonomic units based on homology searches against available annotated data in databases. No reference database covers all biological functions, the ability to visualize and merge the interpretations of all database searches within a single framework is important, as implemented in the MG-RAST and IMG/M servers. IMG/M provides a standardized pipeline, but with higher sensitivity as it performs, for example, hidden Markov model (HMM) and BLASTX searches at substantial computational cost. In contrast to MG-RAST, comparisons in IMG/M are not performed on an abundance table level, but are based on an all vs. all genes comparison. Therefore IMG/M is the only system that integrates all datasets into a single protein level abstraction. Both IMG/M and MG-RAST provide the ability to use stored computational results for comparison, enabling comparison of novel metagenomes with a rich body of other datasets without requiring the end-user to provide the computational means for reanalysis of all datasets involved in their study. (this metagenome/transcriptom annotation section was mainly quoted from metagenomics review article (Thomas, Gilbert et al. 2012).

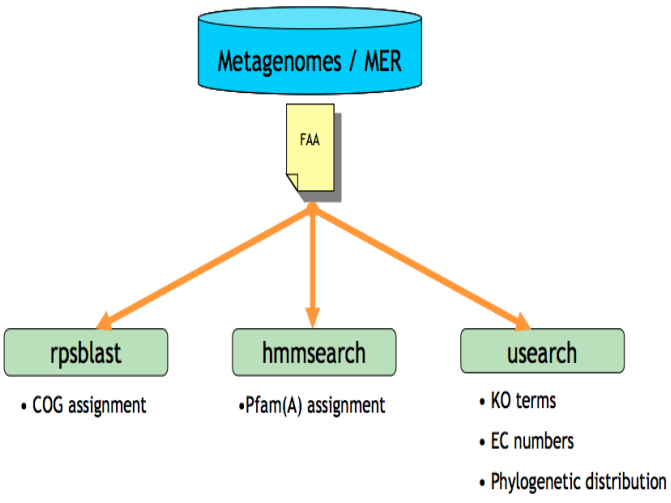

Figure 4. IMG/M annotation pipeline.

Single cell pipeline

Functional based single cell sorting and sequencing

Metagenome assembly and annotation can be improved dramatically with increase of reference genomes of the microbial community, thus functional analysis in metatranscriptomics can gain more resolution of the functions in the community and its players. Single cell sequencing is positioned as an emerging tool to obtain genome sequences from uncultivated microbes and expanded branches of the tree of life in "dark matter microbes"(Rinke, Schwientek et al. 2013). Thus, it has contributed to a more unbiased representation of genes and functions in the references for metagenomics and transcriptomics analysis. Here I propose to incorporate a single cell analysis pipeline as part of the metagenomics pipeline to improve analysis of metatranscriptomic data and to directly connect the functions in the community to the microbes.

In general single cell approaches, single cell genomes are obtained by fluorescent associated cell sorting (FACS) or micromanipulator isolation followed by MDA (multiple displacement amplification) using phi29 polymerase (Dean Fb Fau - Nelson, Nelson Jr Fau - Giesler et al.). Briefly, single cells are deposited, lysed and amplified by MDA in single wells of 384 plates. The MDA positive wells are screened by 16s rRNA DNA sequencing and further whole genome sequencing is done on the single genome of interest based on the 16s identification.

Our group has been working on single cell agarose bead approach, where the single cells are captured in agarose beads using Poisson distribution calculation and beads are pulled together for the downstream process for genome amplification. The beads are served as physical porous containers for individual genomic DNA, where small molecules can be exchanged for different reactions. A distinct advantage of this approach is that it allows us to be able to do fluorescent in-situ hybridization (FISH) on target genes since the beads hold more than ten thousands of fold

amplified genomic DNA. Thus, the whole genome of interest carrying the functional genes can be labeled and sorted out for whole genome sequencing and analysis. These sequenced and annotated single genomes are deposited as references in the custom database and iterate searches with unresolved transcripts. Importantly, it also provides direct connection to the functions and microbes that carry the functions in the community.

Single cells assembly and annotation

Single cell assemblers are designed to handle the MDA related complications in sequencing data, very non-uniform read coverage, elevated sequencing errors and chimeric reads. E+V-SC assembler was introduced by Chitsaz, et al., combining the modified versions of well known genome assemblers, a EULER-SR and Velvet, and achieved a significant improvement in fragment assembly for single cells sequencing data (Chitsaz H Fau - Yee-Greenbaum, Yee-Greenbaum Jl Fau - Tesler et al.). IDBA-UD assembler (Peng, Leung et al. 2012) also is based on genome assembler, IDBA, and is extended to single cell or metagenomic sequencing data, which shares uneven sequencing read coverage, by adapting *de Bruijin graph* approach. The recently introduced SPAdes algorithm is designed for the single cell sequence assembly in the first place, rather than just being a modification of existing tools, to fully utilize the potential of single cell sequencing (Bankevich A Fau - Nurk, Nurk S Fau - Antipov et al.). SPAdes uses paired assembly graphs, adopted from *paired de Bruijin graph* approach (Medvedev et al., 2011), after an error correction procedure. It comes in three separate modules: BayesHammer (Nikolenko, Korobeynikov et al. 2013), read error correction tools, SPAdes, iterative short-read genome assembly module, and MismatchCorrctor which improves mismatch and short indel rates in resulting contigs and scaffolds. So far, SPAdes is limited to prokaryotic genome assembly. However, SPAdes out-performs other existing single cell assemblers for small sized genomes such as prokaryotic genomes and is chosen for the single cell pipeline.

The annotation is performed by Prokka, *Prokaryotic Genome Annotation System (*http://vicbioinformatics.com/*).* Prokka is a very fast annotation pipeline (10minuts on a typical quad-core desktop for annotating 4Mb genome) and is easy to incorporate into a custom-built pipeline. It comes with custom BLAST and HMMER databases and options to easily modify the databases. Also the output file formats include standard GenBank file format and also accepted formats that other genomic tools require, such as giff3 format. Prokka shares many algorithms mentioned in the gene annotation section: Aragorn for tRNA scan, Barrnap (license-free replacement for RNAmmer) for ribosomal RNA scan, Infernal for similarity searching against ncRNA family profiles, Prodigal for coding gene prediction, BLAST+ for similarity searching against protein sequence libraries, and HMMER3 for similarity searching against protein family profiles.

Summary

It is still far distant that we achieve a high-resolution view of complex microbial communities and understand how they interact with their environments. I attempt to advance the understanding of functions and microbes carry the functions in the community by proposing a metatranscriptomics pipeline to link functions to microbes in the community. Mainly by incorporating single cell bead and functional sorting of genome approaches the metatranscriptomics pipeline efficiently increases reference genomes directly related to the functions of interest and enables connection of the functions to the taxa in the community. Many details and decisions for implementing this metatransciptomics pipeline remain unresolved and need to be worked out by more research and by evaluations of existing tools and pipelines.

References

Bankevich A Fau - Nurk, S., et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." (1557-8666 (Electronic)).

The lion's share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies. A major goal of single-cell genomics is to complement gene-centric metagenomic data with whole-genome assemblies of uncultivated organisms. Assembly of single-cell data is challenging because of highly non-uniform read coverage as well as elevated levels of sequencing errors and chimeric reads. We describe SPAdes, a new assembler for both single-cell and standard (multicell) assembly, and demonstrate that it improves on the recently released E+V-SC assembler (specialized for single-cell data) and on popular assemblers Velvet and SoapDeNovo (for multicell data). SPAdes generates single-cell assemblies, providing information about genomes of uncultivatable bacteria that vastly exceeds what may be obtained via traditional metagenomics studies. SPAdes is available online (http://bioinf.spbau.ru/spades). It is distributed as open source software.

Besemer, J., et al. (2001). "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions." Nucleic Acids Res **29**(12): 2607-2618.

Improving the accuracy of prediction of gene starts is one of a few remaining open problems in computer prediction of prokaryotic genes. Its difficulty is caused by the absence of relatively strong sequence patterns identifying true translation initiation sites. In the current paper we show that the accuracy of gene start prediction can be improved by combining models of protein-coding and non-coding regions and models of regulatory sites near gene start within an iterative Hidden Markov model based algorithm. The new gene prediction method, called GeneMarkS, utilizes a non-supervised training procedure and can be used for a newly sequenced prokaryotic genome with no prior knowledge of any protein or rRNA genes. The GeneMarkS implementation uses an improved version of the gene finding program GeneMark.hmm, heuristic Markov models of coding and non-coding regions and the Gibbs sampling multiple alignment program. GeneMarkS predicted precisely 83.2% of the translation starts of GenBank annotated Bacillus subtilis genes and 94.4% of translation starts in an experimentally validated set of Escherichia coli genes. We have also observed that GeneMarkS detects prokaryotic genes, in terms of identifying open reading frames containing real genes, with an accuracy matching the level of the best currently used gene detection methods. Accurate translation start prediction, in addition to the refinement of protein sequence N-terminal data, provides the benefit of precise positioning of the sequence region situated upstream to a gene start. Therefore, sequence motifs related to transcription and translation regulatory sites can be revealed and analyzed with higher precision. These

motifs were shown to possess a significant variability, the functional and evolutionary connections of which are discussed.

Bland, C., et al. (2007). "CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats." BMC Bioinformatics **8**(1): 209.

BACKGROUND:Clustered Regularly Interspaced Palindromic Repeats (CRISPRs) are a novel type of direct repeat found in a wide range of bacteria and archaea. CRISPRs are beginning to attract attention because of their proposed mechanism; that is, defending their hosts against invading extrachromosomal elements such as viruses. Existing repeat detection tools do a poor job of identifying CRISPRs due to the presence of unique spacer sequences separating the repeats. In this study, a new tool, CRT, is introduced that rapidly and accurately identifies CRISPRs in large DNA strings, such as genomes and metagenomes.RESULTS:CRT was compared to CRISPR detection tools, Patscan and Pilercr. In terms of correctness, CRT was shown to be very reliable, demonstrating significant improvements over Patscan for measures precision, recall and quality. When compared to Pilercr, CRT showed improved performance for recall and quality. In terms of speed, CRT proved to be a huge improvement over Patscan. Both CRT and Pilercr were comparable in speed, however CRT was faster for genomes containing large numbers of repeats.CONCLUSION:In this paper a new tool was introduced for the automatic detection of CRISPR elements. This tool, CRT, showed some important improvements over current techniques for CRISPR identification. CRT's approach to detecting repetitive sequences is straightforward. It uses a simple sequential scan of a DNA sequence and detects repeats directly without any major conversion or preprocessing of the input. This leads to a program that is easy to describe and understand; yet it is very accurate, fast and memory efficient, being O(n) in space and O(nm/l) in time.

Caporaso, J. G., et al. (2010). "QIIME allows analysis of high-throughput community sequencing data." Nat Methods **7**(5): 335-336.

Caporaso, J. G., et al. (2010). "QIIME allows analysis of high-throughput community sequencing data." Nat Meth **7**(5): 335-336.

Caporaso, J. G., et al. (2012). "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms." ISME J **6**(8): 1621-1624.

Chitsaz H Fau - Yee-Greenbaum, J. L., et al. "Efficient de novo assembly of single-cell bacterial genomes from short-read data sets." (1546-1696 (Electronic)).

Whole genome amplification by the multiple displacement amplification (MDA) method allows sequencing of DNA from single cells of bacteria that cannot be cultured. Assembling a genome is challenging, however, because MDA generates highly nonuniform coverage of the genome. Here we describe

an algorithm tailored for short-read data from single cells that improves assembly through the use of a progressively increasing coverage cutoff. Assembly of reads from single Escherichia coli and Staphylococcus aureus cells captures >91% of genes within contigs, approaching the 95% captured from an assembly based on many E. coli cells. We apply this method to assemble a genome from a single cell of an uncultivated SAR324 clade of Deltaproteobacteria, a cosmopolitan bacterial lineage in the global ocean. Metabolic reconstruction suggests that SAR324 is aerobic, motile and chemotaxic. Our approach enables acquisition of genome assemblies for individual uncultivated bacteria using only short reads, providing cell-specific genetic information absent from metagenomic studies.

Dean Fb Fau - Nelson, J. R., et al. "Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification." (1088-9051 (Print)).

We describe a simple method of using rolling circle amplification to amplify vector DNA such as M13 or plasmid DNA from single colonies or plaques. Using random primers and phi29 DNA polymerase, circular DNA templates can be amplified 10,000-fold in a few hours. This procedure removes the need for lengthy growth periods and traditional DNA isolation methods. Reaction products can be used directly for DNA sequencing after phosphatase treatment to inactivate unincorporated nucleotides. Amplified products can also be used for in vitro cloning, library construction, and other molecular biology applications.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.

The recent literature on profile hidden Markov model (profile HMM) methods and software is reviewed. Profile HMMs turn a multiple sequence alignment into a position-specific scoring system suitable for searching databases for remotely homologous sequences. Profile HMM analyses complement standard pairwise comparison methods for large-scale sequence analysis. Several software implementations and two large libraries of profile HMMs of common protein domains are available. HMM methods performed comparably to threading methods in the CASP2 structure prediction exercise.

Falgueras J Fau - Lara, A. J., et al. "SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read." (1471-2105 (Electronic)).

BACKGROUND: High-throughput automated sequencing has enabled an exponential growth rate of sequencing data. This requires increasing sequence quality and reliability in order to avoid database contamination with artefactual sequences. The arrival of pyrosequencing enhances this problem and necessitates customisable pre-processing algorithms. RESULTS: SeqTrim has been implemented both as a Web and as a standalone command line application. Already-published and newly-designed algorithms have been included to identify sequence inserts, to remove low quality, vector,

adaptor, low complexity and contaminant sequences, and to detect chimeric reads. The availability of several input and output formats allows its inclusion in sequence processing workflows. Due to its specific algorithms, SeqTrim outperforms other pre-processors implemented as Web services or standalone applications. It performs equally well with sequences from EST libraries, SSH libraries, genomic DNA libraries and pyrosequencing reads and does not lead to over-trimming. CONCLUSIONS: SeqTrim is an efficient pipeline designed for pre-processing of any type of sequence read, including next-generation sequencing. It is easily configurable and provides a friendly interface that allows users to know what happened with sequences at every pre-processing stage, and to verify pre-processing of an individual sequence if desired. The recommended pipeline reveals more information about each sequence than previously described pre-processors and can discard more sequencing or experimental artefacts.

Griffiths-Jones, S., et al. (2005). "Rfam: annotating non-coding RNAs in complete genomes." Nucleic Acids Res **33**(suppl 1): D121-D124.

Rfam is a comprehensive collection of non-coding RNA (ncRNA) families, represented by multiple sequence alignments and profile stochastic context-free grammars. Rfam aims to facilitate the identification and classification of new members of known sequence families, and distributes annotation of ncRNAs in over 200 complete genome sequences. The data provide the first glimpses of conservation of multiple ncRNA families across a wide taxonomic range. A small number of large families are essential in all three kingdoms of life, with large numbers of smaller families specific to certain taxa. Recent improvements in the database are discussed, together with challenges for the future. Rfam is available on the Web at http://www.sanger.ac.uk/Software/Rfam/ and http://rfam.wustl.edu/.

Haft, D. H., et al. (2003). "The TIGRFAMs database of protein families." Nucleic Acids Res **31**(1): 371-373.

TIGRFAMs is a collection of manually curated protein families consisting of hidden Markov models (HMMs), multiple sequence alignments, commentary, Gene Ontology (GO) assignments, literature references and pointers to related TIGRFAMs, Pfam and InterPro models. These models are designed to support both automated and manually curated annotation of genomes. TIGRFAMs contains models of full-length proteins and shorter regions at the levels of superfamilies, subfamilies and equivalogs, where equivalogs are sets of homologous proteins conserved with respect to function since their last common ancestor. The scope of each model is set by raising or lowering cutoff scores and choosing members of the seed alignment to group proteins sharing specific function (equivalog) or more general properties. The overall goal is to provide information with maximum utility for the annotation process. TIGRFAMs is thus complementary to Pfam, whose models typically achieve broad coverage across distant homologs but end at the boundaries of conserved structural domains. The database currently contains over 1600

protein families. TIGRFAMs is available for searching or downloading at http://www.tigr.org/TIGRFAMs.

He, S., et al. (2010). "Validation of two ribosomal RNA removal methods for microbial metatranscriptomics." Nat Meth **7**(10): 807-812.

Hoff, K. J., et al. (2009). "Orphelia: predicting genes in metagenomic sequencing reads." Nucleic Acids Res **37**(suppl 2): W101-W105.
Metagenomic sequencing projects yield numerous sequencing reads of a diverse range of uncultivated and mostly yet unknown microorganisms. In many cases, these sequencing reads cannot be assembled into longer contigs. Thus, gene prediction tools that were originally developed for whole-genome analysis are not suitable for processing metagenomes. Orphelia is a program for predicting genes in short DNA sequences that is available through a web server application (http://orphelia.gobics.de). Orphelia utilizes prediction models that were created with machine learning techniques on the basis of a wide range of annotated genomes. In contrast to other methods for metagenomic gene prediction, Orphelia has fragment length-specific prediction models for the two most popular sequencing techniques in metagenomics, chain termination sequencing and pyrosequencing. These models ensure highly specific gene predictions.

Hyatt, D., et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC Bioinformatics **11**(1): 119.
BACKGROUND:The quality of automated gene prediction in microbial organisms has improved steadily over the past decade, but there is still room for improvement. Increasing the number of correct identifications, both of genes and of the translation initiation sites for each gene, and reducing the overall number of false positives, are all desirable goals.RESULTS:With our years of experience in manually curating genomes for the Joint Genome Institute, we developed a new gene prediction algorithm called Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm). With Prodigal, we focused specifically on the three goals of improved gene structure prediction, improved translation initiation site recognition, and reduced false positives. We compared the results of Prodigal to existing gene-finding methods to demonstrate that it met each of these objectives.CONCLUSION:We built a fast, lightweight, open source gene prediction program called Prodigal http://compbio.ornl.gov/prodigal/. Prodigal achieved good results compared to existing methods, and we believe it will be a valuable asset to automated microbial annotation pipelines.

Kanehisa, M., et al. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(suppl 1): D480-D484.
KEGG (http://www.genome.jp/kegg/) is a database of biological systems that integrates genomic, chemical and systemic functional information. KEGG

provides a reference knowledge base for linking genomes to life through the process of PATHWAY mapping, which is to map, for example, a genomic or transcriptomic content of genes to KEGG reference pathways to infer systemic behaviors of the cell or the organism. In addition, KEGG provides a reference knowledge base for linking genomes to the environment, such as for the analysis of drug-target relationships, through the process of BRITE mapping. KEGG BRITE is an ontology database representing functional hierarchies of various biological objects, including molecules, cells, organisms, diseases and drugs, as well as relationships among them. KEGG PATHWAY is now supplemented with a new global map of metabolic pathways, which is essentially a combined map of about 120 existing pathway maps. In addition, smaller pathway modules are defined and stored in KEGG MODULE that also contains other functional units and complexes. The KEGG resource is being expanded to suit the needs for practical applications. KEGG DRUG contains all approved drugs in the US and Japan, and KEGG DISEASE is a new database linking disease genes, pathways, drugs and diagnostic markers.

Kuczynski, J., et al. (2012). "Using QIIME to analyze 16S rRNA gene sequences from microbial communities." Curr Protoc Microbiol **Chapter 1**: Unit 1E 5.

QIIME (canonically pronounced "chime") is a software application that performs microbial community analysis. It is an acronym for Quantitative Insights Into Microbial Ecology, and has been used to analyze and interpret nucleic acid sequence data from fungal, viral, bacterial, and archaeal communities. The following protocols describe how to install QIIME on a single computer and use it to analyze microbial 16S sequence data from nine distinct microbial communities.

Lagesen, K., et al. (2007). "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." Nucleic Acids Res **35**(9): 3100-3108.

The publication of a complete genome sequence is usually accompanied by annotations of its genes. In contrast to protein coding genes, genes for ribosomal RNA (rRNA) are often poorly or inconsistently annotated. This makes comparative studies based on rRNA genes difficult. We have therefore created computational predictors for the major rRNA species from all kingdoms of life and compiled them into a program called RNAmmer. The program uses hidden Markov models trained on data from the 5S ribosomal RNA database and the European ribosomal RNA database project. A pre-screening step makes the method fast with little loss of sensitivity, enabling the analysis of a complete bacterial genome in less than a minute. Results from running RNAmmer on a large set of genomes indicate that the location of rRNAs can be predicted with a very high level of accuracy. Novel, unannotated rRNAs are also predicted in many genomes. The software as well as the genome analysis results are available at the CBS web server.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Meth **9**(4): 357-359.

Laslett, D. and B. Canback (2004). "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences." Nucleic Acids Res **32**(1): 11-16.

A computer program, ARAGORN, identifies tRNA and tmRNA genes. The program employs heuristic algorithms to predict tRNA secondary structure, based on homology with recognized tRNA consensus sequences and ability to form a base‐paired cloverleaf. tmRNA genes are identified using a modified version of the BRUCE program. ARAGORN achieves a detection sensitivity of 99% from a set of 1290 eubacterial, eukaryotic and archaeal tRNA genes and detects all complete tmRNA sequences in the tmRNA database, improving on the performance of the BRUCE program. Recently discovered tmRNA genes in the chloroplasts of two species from the 'green' algae lineage are detected. The output of the program reports the proposed tRNA secondary structure and, for tmRNA genes, the secondary structure of the tRNA domain, the tmRNA gene sequence, the tag peptide and a list of organisms with matching tmRNA peptide tags.

Lassmann T Fau - Hayashizaki, Y., et al. "TagDust--a program to eliminate artifacts from next generation sequencing data." (1367-4811 (Electronic)).

MOTIVATION: Next-generation parallel sequencing technologies produce large quantities of short sequence reads. Due to experimental procedures various types of artifacts are commonly sequenced alongside the targeted RNA or DNA sequences. Identification of such artifacts is important during the development of novel sequencing assays and for the downstream analysis of the sequenced libraries. RESULTS: Here we present TagDust, a program identifying artifactual sequences in large sequencing runs. Given a user-defined cutoff for the false discovery rate, TagDust identifies all reads explainable by combinations and partial matches to known sequences used during library preparation. We demonstrate the quality of our method on sequencing runs performed on Illumina's Genome Analyzer platform. AVAILABILITY: Executables and documentation are available from http://genome.gsc.riken.jp/osc/english/software/. CONTACT: timolassmann@gmail.com.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of

MAQ is also a concern when the alignment is scaled up to the resequencing of hundreds of individuals.Results: We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with Burrows–Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. BWA supports both base space reads, e.g. from Illumina sequencing machines, and color space reads from AB SOLiD machines. Evaluations on both simulated and real data suggest that BWA is ~10–20× faster than MAQ, while achieving similar accuracy. In addition, BWA outputs alignment in the new standard SAM (Sequence Alignment/Map) format. Variant calling and other downstream analyses after the alignment can be achieved with the open source SAMtools software package.Availability: http://maq.sourceforge.netContact: rd@sanger.ac.uk

Li, R., et al. (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**(5): 713-714.

Summary: We have developed a program SOAP for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The program is designed to handle the huge amounts of short reads generated by parallel sequencing using the new generation Illumina-Solexa sequencing technology. SOAP is compatible with numerous applications, including single-read or pair-end resequencing, small RNA discovery and mRNA tag sequence mapping. SOAP is a command-driven program, which supports multi-threaded parallel computing, and has a batch module for multiple query sets.Availability: http://soap.genomics.org.cnContact: soap@genomics.org.cn

Li, R., et al. (2009). "SOAP2: an improved ultrafast tool for short read alignment." Bioinformatics **25**(15): 1966-1967.

Summary: SOAP2 is a significantly improved version of the short oligonucleotide alignment program that both reduces computer memory usage and increases alignment speed at an unprecedented rate. We used a Burrows Wheeler Transformation (BWT) compression index to substitute the seed strategy for indexing the reference sequence in the main memory. We tested it on the whole human genome and found that this new algorithm reduced memory usage from 14.7 to 5.4 GB and improved alignment speed by 20–30 times. SOAP2 is compatible with both single- and paired-end reads. Additionally, this tool now supports multiple text and compressed file formats. A consensus builder has also been developed for consensus assembly and SNP detection from alignment of short reads on a reference genome.Availability: http://soap.genomics.org.cnContact: soap@genomics.org.cn

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence." Nucleic Acids Res **25**(5): 0955-0964.

We describe a program, tRNAscan-SE, which identifies 99–100% of transfer RNA genes in DNA sequence while giving less than one false positive per 15 gigabases. Two previously described tRNA detection programs are used as fast, first-pass prefilters to identify candidate tRNAs, which are then analyzed by a highly selective tRNA covariance model. This work represents a practical application of RNA covariance models, which are general, probabilistic secondary structure profiles based on stochastic context-free grammars. tRNAscan-SE searches at ~30 000 bp/s. Additional extensions to tRNAscan-SE detect unusual tRNA homologues such as selenocysteine tRNAs, tRNA-derived repetitive elements and tRNA pseudo-genes.

Lunter, G. and M. Goodson (2011). "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads." Genome Research **21**(6): 936-939.

High-volume sequencing of DNA and RNA is now within reach of any research laboratory and is quickly becoming established as a key research tool. In many workflows, each of the short sequences ("reads") resulting from a sequencing run are first "mapped" (aligned) to a reference sequence to infer the read from which the genomic location derived, a challenging task because of the high data volumes and often large genomes. Existing read mapping software excel in either speed (e.g., BWA, Bowtie, ELAND) or sensitivity (e.g., Novoalign), but not in both. In addition, performance often deteriorates in the presence of sequence variation, particularly so for short insertions and deletions (indels). Here, we present a read mapper, Stampy, which uses a hybrid mapping algorithm and a detailed statistical model to achieve both speed and sensitivity, particularly when reads include sequence variation. This results in a higher useable sequence yield and improved accuracy compared to that of existing software.

Marchler-Bauer, A., et al. (2007). "CDD: a conserved domain database for interactive domain family analysis." Nucleic Acids Res **35**(suppl 1): D237-D240.

The conserved domain database (CDD) is part of NCBI's Entrez database system and serves as a primary resource for the annotation of conserved domain footprints on protein sequences in Entrez. Entrez's global query interface can be accessed at http://www.ncbi.nlm.nih.gov/Entrez and will search CDD and many other databases. Domain annotation for proteins in Entrez has been pre-computed and is readily available in the form of 'Conserved Domain' links. Novel protein sequences can be scanned against CDD using the CD-Search service; this service searches databases of CDD-derived profile models with protein sequence queries using BLAST heuristics, at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. Protein query sequences submitted to NCBI's protein BLAST search service are scanned for conserved domain signatures by default. The CDD collection contains models imported from Pfam, SMART and COG, as well as domain models curated at NCBI. NCBI curated models are organized into hierarchies of domains related by common descent. Here we report on the status of the curation effort and present a novel helper application, CDTree, which enables

users of the CDD resource to examine curated hierarchies. More importantly, CDD and CDTree used in concert, serve as a powerful tool in protein classification, as they allow users to analyze protein sequences in the context of domain family hierarchies.

Markowitz, V. M., et al. (2012). "IMG/M: the integrated metagenome data management and comparative analysis system." Nucleic Acids Res **40**(D1): D123-D129.

The integrated microbial genomes and metagenomes (IMG/M) system provides support for comparative analysis of microbial community aggregate genomes (metagenomes) in a comprehensive integrated context. IMG/M integrates metagenome data sets with isolate microbial genomes from the IMG system. IMG/M's data content and analytical capabilities have been extended through regular updates since its first release in 2007. IMG/M is available at http://img.jgi.doe.gov/m. A companion IMG/M systems provide support for annotation and expert review of unpublished metagenomic data sets (IMG/M ER: http://img.jgi.doe.gov/mer).

Martin, J., et al. (2010). "Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads." BMC Genomics **11**: 663.

BACKGROUND: Comprehensive annotation and quantification of transcriptomes are outstanding problems in functional genomics. While high throughput mRNA sequencing (RNA-Seq) has emerged as a powerful tool for addressing these problems, its success is dependent upon the availability and quality of reference genome sequences, thus limiting the organisms to which it can be applied. RESULTS: Here, we describe Rnnotator, an automated software pipeline that generates transcript models by de novo assembly of RNA-Seq data without the need for a reference genome. We have applied the Rnnotator assembly pipeline to two yeast transcriptomes and compared the results to the reference gene catalogs of these organisms. The contigs produced by Rnnotator are highly accurate (95%) and reconstruct full-length genes for the majority of the existing gene models (54.3%). Furthermore, our analyses revealed many novel transcribed regions that are absent from well annotated genomes, suggesting Rnnotator serves as a complementary approach to analysis based on a reference genome for comprehensive transcriptomics. CONCLUSIONS: These results demonstrate that the Rnnotator pipeline is able to reconstruct full-length transcripts in the absence of a complete reference genome.

Martin, J. A. and Z. Wang (2011). "Next-generation transcriptome assembly." Nat Rev Genet **12**(10): 671-682.

Transcriptomics studies often rely on partial reference transcriptomes that fail to capture the full catalogue of transcripts and their variations. Recent advances in sequencing technologies and assembly algorithms have facilitated the reconstruction of the entire transcriptome by deep RNA sequencing (RNA-seq), even without a reference genome. However,

transcriptome assembly from billions of RNA-seq reads, which are often very short, poses a significant informatics challenge. This Review summarizes the recent developments in transcriptome assembly approaches - reference-based, de novo and combined strategies - along with some perspectives on transcriptome assembly in the near future.

Nawrocki, E. P., et al. (2009). "Infernal 1.0: inference of RNA alignments." Bioinformatics **25**(10): 1335-1337.

Summary: infernal builds consensus RNA secondary structure profiles called covariance models (CMs), and uses them to search nucleic acid sequence databases for homologous RNAs, or to create new sequence- and structure-based multiple sequence alignments.Availability: Source code, documentation and benchmark downloadable from http://infernal.janelia.org. infernal is freely licensed under the GNU GPLv3 and should be portable to any POSIX-compliant operating system, including Linux and Mac OS/X.Contact: nawrockie,kolbed,eddys@janelia.hhmi.org

Nikolenko, S., et al. (2013). "BayesHammer: Bayesian clustering for error correction in single-cell sequencing." BMC Genomics **14**(Suppl 1): S7.

Error correction of sequenced reads remains a difficult task, especially in single-cell sequencing projects with extremely non-uniform coverage. While existing error correction tools designed for standard (multi-cell) sequencing data usually come up short in single-cell sequencing projects, algorithms actually used for single-cell error correction have been so far very simplistic.We introduce several novel algorithms based on Hamming graphs and Bayesian subclustering in our new error correction tool BAYESHAMMER. While BAYESHAMMER was designed for single-cell sequencing, we demonstrate that it also improves on existing error correction tools for multi-cell sequencing data while working much faster on real-life datasets. We benchmark BAYESHAMMER on both k-mer counts and actual assembly results with the SPADES genome assembler.

Noguchi, H., et al. (2006). "MetaGene: prokaryotic gene finding from environmental genome shotgun sequences." Nucleic Acids Res **34**(19): 5623-5630.

Exhaustive gene identification is a fundamental goal in all metagenomics projects. However, most metagenomic sequences are unassembled anonymous fragments, and conventional gene-finding methods cannot be applied. We have developed a prokaryotic gene-finding program, MetaGene, which utilizes di-codon frequencies estimated by the GC content of a given sequence with other various measures. MetaGene can predict a whole range of prokaryotic genes based on the anonymous genomic sequences of a few hundred bases, with a sensitivity of 95% and a specificity of 90% for artificial shotgun sequences (700 bp fragments from 12 species). MetaGene has two sets of codon frequency interpolations, one for bacteria and one for archaea, and automatically selects the proper set for a given sequence using the domain classification method we propose. The domain classification works

properly, correctly assigning domain information to more than 90% of the artificial shotgun sequences. Applied to the Sargasso Sea dataset, MetaGene predicted almost all of the annotated genes and a notable number of novel genes. MetaGene can be applied to wide variety of metagenomic projects and expands the utility of metagenomics.

Peng, Y., et al. (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics **28**(11): 1420-1428.

Motivation: Next-generation sequencing allows us to sequence reads from a microbial environment using single-cell sequencing or metagenomic sequencing technologies. However, both technologies suffer from the problem that sequencing depth of different regions of a genome or genomes from different species are highly uneven. Most existing genome assemblers usually have an assumption that sequencing depths are even. These assemblers fail to construct correct long contigs.Results: We introduce the IDBA-UD algorithm that is based on the de Bruijn graph approach for assembling reads from single-cell sequencing or metagenomic sequencing technologies with uneven sequencing depths. Several non-trivial techniques have been employed to tackle the problems. Instead of using a simple threshold, we use multiple depthrelative thresholds to remove erroneous k-mers in both low-depth and high-depth regions. The technique of local assembly with paired-end information is used to solve the branch problem of low-depth short repeat regions. To speed up the process, an error correction step is conducted to correct reads of high-depth regions that can be aligned to highconfident contigs. Comparison of the performances of IDBA-UD and existing assemblers (Velvet, Velvet-SC, SOAPdenovo and Meta-IDBA) for different datasets, shows that IDBA-UD can reconstruct longer contigs with higher accuracy.Availability: The IDBA-UD toolkit is available at our website http://www.cs.hku.hk/~alse/idba_udContact: chin@cs.hku.hk

Punta, M., et al. (2012). "The Pfam protein families database." Nucleic Acids Res **40**(D1): D290-D301.

Pfam is a widely used database of protein families, currently containing more than 13 000 manually curated protein families as of release 26.0. Pfam is available via servers in the UK (http://pfam.sanger.ac.uk/), the USA (http://pfam.janelia.org/) and Sweden (http://pfam.sbc.su.se/). Here, we report on changes that have occurred since our 2010 NAR paper (release 24.0). Over the last 2 years, we have generated 1840 new families and increased coverage of the UniProt Knowledgebase (UniProtKB) to nearly 80%. Notably, we have taken the step of opening up the annotation of our families to the Wikipedia community, by linking Pfam families to relevant Wikipedia pages and encouraging the Pfam and Wikipedia communities to improve and expand those pages. We continue to improve the Pfam website and add new visualizations, such as the 'sunburst' representation of taxonomic distribution of families. In this work we additionally address two

topics that will be of particular interest to the Pfam community. First, we explain the definition and use of family-specific, manually curated gathering thresholds. Second, we discuss some of the features of domains of unknown function (also known as DUFs), which constitute a rapidly growing class of families within Pfam.

Rho, M., et al. (2010). "FragGeneScan: predicting genes in short and error-prone reads." Nucleic Acids Res **38**(20): e191.

The advances of next-generation sequencing technology have facilitated metagenomics research that attempts to determine directly the whole collection of genetic material within an environmental sample (i.e. the metagenome). Identification of genes directly from short reads has become an important yet challenging problem in annotating metagenomes, since the assembly of metagenomes is often not available. Gene predictors developed for whole genomes (e.g. Glimmer) and recently developed for metagenomic sequences (e.g. MetaGene) show a significant decrease in performance as the sequencing error rates increase, or as reads get shorter. We have developed a novel gene prediction method FragGeneScan, which combines sequencing error models and codon usages in a hidden Markov model to improve the prediction of protein-coding region in short reads. The performance of FragGeneScan was comparable to Glimmer and MetaGene for complete genomes. But for short reads, FragGeneScan consistently outperformed MetaGene (accuracy improved ~62% for reads of 400 bases with 1% sequencing errors, and ~18% for short reads of 100 bases that are error free). When applied to metagenomes, FragGeneScan recovered substantially more genes than MetaGene predicted (>90% of the genes identified by homology search), and many novel genes with no homologs in current protein sequence database.

Rinke, C., et al. (2013). "Insights into the phylogeny and coding potential of microbial dark matter." Nature **499**(7459): 431-437.

Genome sequencing enhances our understanding of the biological world by providing blueprints for the evolutionary and functional diversity that shapes the biosphere. However, microbial genomes that are currently available are of limited phylogenetic breadth, owing to our historical inability to cultivate most microorganisms in the laboratory. We apply single-cell genomics to target and sequence 201 uncultivated archaeal and bacterial cells from nine diverse habitats belonging to 29 major mostly uncharted branches of the tree of life, so-called 'microbial dark matter'. With this additional genomic information, we are able to resolve many intra- and inter-phylum-level relationships and to propose two new superphyla. We uncover unexpected metabolic features that extend our understanding of biology and challenge established boundaries between the three domains of life. These include a novel amino acid use for the opal stop codon, an archaeal-type purine synthesis in Bacteria and complete sigma factors in Archaea similar to those in Bacteria. The single-cell genomes also served to phylogenetically

anchor up to 20% of metagenomic reads in some habitats, facilitating organism-level interpretation of ecosystem function. This study greatly expands the genomic representation of the tree of life and provides a systematic step towards a better understanding of biological evolution on our planet.

Salzberg, S. L., et al. (1998). "Microbial gene identification using interpolated Markov models." Nucleic Acids Res **26**(2): 544-548.

This paper describes a new system, GLIMMER, for finding genes in microbial genomes. In a series of tests on Haemophilus influenzae, Helicobacter pylori and other complete microbial genomes, this system has proven to be very accurate at locating virtually all the genes in these sequences, outperforming previous methods. A conservative estimate based on experiments on H.pylori and H.influenzae is that the system finds >97% of all genes. GLIMMER uses interpolated Markov models (IMMs) as a framework for capturing dependencies between nearby nucleotides in a DNA sequence. An IMM-based method makes predictions based on a variable context; i.e., a variable-length oligomer in a DNA sequence. The context used by GLIMMER changes depending on the local composition of the sequence. As a result, GLIMMER is more flexible and more powerful than fixed-order Markov methods, which have previously been the primary content-based technique for finding genes in microbial DNA.

Smith, D. R., et al. (2008). "Rapid whole-genome mutational profiling using next-generation sequencing technologies." Genome Research **18**(10): 1638-1642.

Forward genetic mutational studies, adaptive evolution, and phenotypic screening are powerful tools for creating new variant organisms with desirable traits. However, mutations generated in the process cannot be easily identified with traditional genetic tools. We show that new high-throughput, massively parallel sequencing technologies can completely and accurately characterize a mutant genome relative to a previously sequenced parental (reference) strain. We studied a mutant strain of Pichia stipitis, a yeast capable of converting xylose to ethanol. This unusually efficient mutant strain was developed through repeated rounds of chemical mutagenesis, strain selection, transformation, and genetic manipulation over a period of seven years. We resequenced this strain on three different sequencing platforms. Surprisingly, we found fewer than a dozen mutations in open reading frames. All three sequencing technologies were able to identify each single nucleotide mutation given at least 10–15-fold nominal sequence coverage. Our results show that detecting mutations in evolved and engineered organisms is rapid and cost-effective at the whole-genome level using new sequencing technologies. Identification of specific mutations in strains with altered phenotypes will add insight into specific gene functions and guide further metabolic engineering efforts.

Stewart, F. J., et al. (2010). "Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics." ISME J **4**(7): 896-907.

Thomas, T., et al. (2012). "Metagenomics - a guide from sampling to data analysis." Microbial Informatics and Experimentation **2**(1): 3.

Metagenomics applies a suite of genomic technologies and bioinformatics tools to directly access the genetic content of entire communities of organisms. The field of metagenomics has been responsible for substantial advances in microbial ecology, evolution, and diversity over the past 5 to 10 years, and many research laboratories are actively engaged in it now. With the growing numbers of activities also comes a plethora of methodological knowledge and expertise that should guide future developments in the field. This review summarizes the current opinions in metagenomics, and provides practical guidance and advice on sample processing, sequencing technology, assembly, binning, annotation, experimental design, statistical analysis, data storage, and data sharing. As more metagenomic datasets are generated, the availability of standardized procedures and shared data storage and analysis becomes increasingly important to ensure that output of individual projects can be assessed and compared.

Zhao, Q. Y., et al. (2011). "Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study." BMC Bioinformatics **12 Suppl 14**: S2.

BACKGROUND: With the fast advances in nextgen sequencing technology, high-throughput RNA sequencing has emerged as a powerful and cost-effective way for transcriptome study. De novo assembly of transcripts provides an important solution to transcriptome analysis for organisms with no reference genome. However, there lacked understanding on how the different variables affected assembly outcomes, and there was no consensus on how to approach an optimal solution by selecting software tool and suitable strategy based on the properties of RNA-Seq data. RESULTS: To reveal the performance of different programs for transcriptome assembly, this work analyzed some important factors, including k-mer values, genome complexity, coverage depth, directional reads, etc. Seven program conditions, four single k-mer assemblers (SK: SOAPdenovo, ABySS, Oases and Trinity) and three multiple k-mer methods (MK: SOAPdenovo-MK, trans-ABySS and Oases-MK) were tested. While small and large k-mer values performed better for reconstructing lowly and highly expressed transcripts, respectively, MK strategy worked well for almost all ranges of expression quintiles. Among SK tools, Trinity performed well across various conditions but took the longest running time. Oases consumed the most memory whereas SOAPdenovo required the shortest runtime but worked poorly to reconstruct full-length CDS. ABySS showed some good balance between resource usage and quality of assemblies. CONCLUSIONS: Our work compared the performance of publicly available transcriptome assemblers, and analyzed important factors affecting de novo assembly. Some practical guidelines for transcript

reconstruction from short-read RNA-Seq data were proposed. De novo assembly of C. sinensis transcriptome was greatly improved using some optimized methods.

Prokka: Prokaryotic Genome Annotation System - http://vicbioinformatics.com/