 doi:10.1093/bib/bbs070

Bioinformatic approaches for functional annotation and pathway inference in metagenomics data

Carlotta De Filippo, Matteo Ramazzotti, Paolo Fontana and Duccio Cavalieri

Submitted: 19th March 2012; Received (in revised form): 5th October 2012

Abstract

Metagenomic approaches are increasingly recognized as a baseline for understanding the ecology and evolution of microbial ecosystems. The development of methods for pathway inference from metagenomics data is of paramount importance to link a phenotype to a cascade of events stemming from a series of connected sets of genes or proteins. Biochemical and regulatory pathways have until recently been thought and modelled within one cell type, one organism, one species. This vision is being dramatically changed by the advent of whole microbiome sequencing studies, revealing the role of symbiotic microbial populations in fundamental biochemical functions. The new landscape we face requires a clear picture of the potentialities of existing tools and development of new tools to characterize, reconstruct and model biochemical and regulatory pathways as the result of integration of function in complex symbiotic interactions of ontologically and evolutionary distinct cell types.

Keywords: *metagenomics; next-generation sequencing; microbiome; pathway analysis; gene annotation*

NEXT-GENERATION SEQUENCING AS CRADLE AND STAGE OF THE METAGENOMICS REVOLUTION

Microbial communities comprise combinations of bacteria, archaea, fungi, yeasts, eukaryotes and viruses, often co-occurring in a single habitat. Until recently, the tools to systematically study global community function and environment at the molecular level were not available, because complex microbial communities are generally not amenable to laboratory study [1].

At the beginning of this century, cultivation-independent diversity studies were limited by the costs and complexity of Sanger-sequencing methods. In the past 10 years, the picture of microbial communities has rapidly passed from black and white to a surprising explosion of bright colours, thanks to the application of next-generation sequencing (NGS) technologies to sequencing of environmental samples (i.e. metagenomics) [2].

Metagenomic approaches allowed the first large-scale insights into the function of complex microbial communities and are increasingly recognized as a baseline for understanding the ecology and evolution of microbial ecosystems as genetic and metabolic networks. The reasonable conclusion is that the entire and fascinating diversity of biosphere cannot be appreciated unless framed in the appropriate meta-context.

This review will focus on the bioinformatics procedures available for functional annotation and pathway inference from metagenomics sequence information (Figure 1 and Table 1). We will initially discuss methods using 16S rRNA genes to derive

Corresponding author. Duccio Cavalieri, Fondazione Edmund Mach, Research and Innovation Centre, Via E. Mach, 1, 38010 San Michele all'Adige, Trento, Italy. Tel.: +39-0461-615153; Fax +39-0461-650956; E-mail: duccio.cavalieri@fmach.it; duccio.cavalieri@unifi.it

Carlotta De Filippo is a molecular biologist with a PhD in pharmacology. Currently, she is a researcher at CRI-FEM (Trento, Italy), studying the role of diet, ethnicity and environment in shaping gut microbiota in human populations.

Matteo Ramazzotti is a biologist with PhD in biochemistry, aggregate professor of bioinformatics at the University of Florence, Italy. He works on metagenomics and microbial genomics.

Paolo Fontana currently is a researcher at CRI-FEM (Trento, Italy), working on assembly and annotation of NGS data.

Duccio Cavalieri has a PhD in Genetics, professor of microbiology at the University of Florence, Italy, and coordinator of the Centre for Computational Biology at CRI-FEM (Trento, Italy).

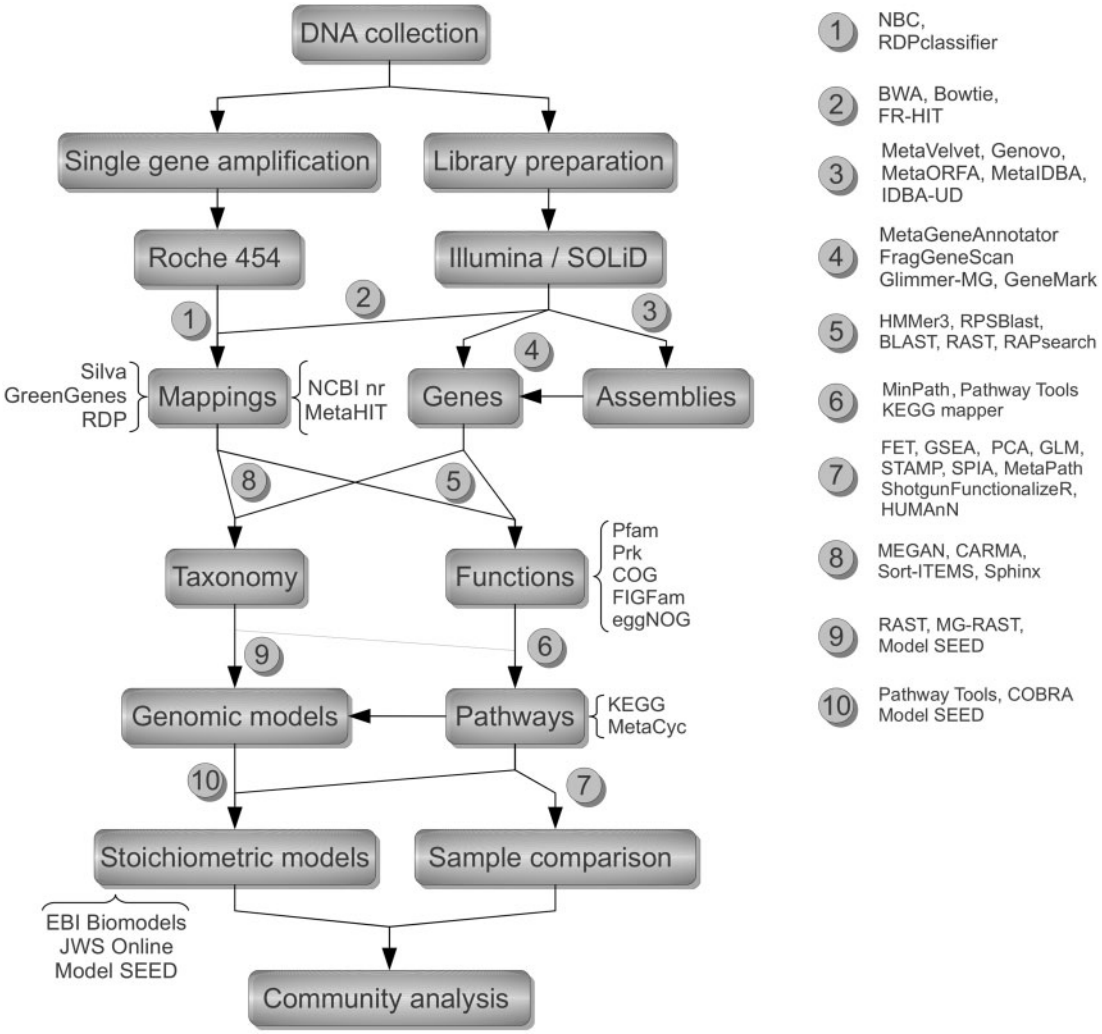

Figure 1: Flowchart of the main steps and bioinformatics tools required for pathway reconstruction from metagenomics surveys. Numbers in circles correspond to specific tools and programs developed for the corresponding steps and listed in the right part of the figure (links listed on Table 1). Curly brackets point to application specific databanks. The analytic procedure ideally bifurcate at the starting point according to the investigation strategy: DNA can undergo a PCR-based amplification step to increase the amount of a specific marker gene (e.g. ribosomal RNA) and then subject to Roche 454 sequencing or can be fragmented and prepared into libraries for metagenomics Illumina/SOLiD sequencing. Both those techniques are characterized by the generation of a huge amount of short reads that necessitate care and powerful instrumentation for their handling and processing. The simplest analytic choice is to map short reads into reference databases such as that maintained by the Ribosomal Database Project for the taxonomy survey via 16 S sequencing (1) or into NCBI non-redundant (nr/nt) for environmental microbiome or, in case of gut microbiome surveys, the better-scoped MetaHIT (2). Another possibility is to assemble the short reads into longer contigs using new generation assemblers specific for unevenly distributed reads deriving from the multitude of different microbes represented in the community (3). Their application improves the efficiency of gene finding programs that, even though applicable directly on reads, have a higher level of information to ensure more confident gene identification (4). Once coding sequences have been obtained, their corresponding proteins can be searched in reference functional databases encoding information in the form of HMMs or PSSM from multiple sequence alignments (5) or directly in reference protein sets derived from primary databanks or from genome-derived collections. The first approach leads to a direct identification of associated functions that can be used to identify and score pathways (6) and in the end apply a battery of statistical techniques for sample characterization (7). The second approach can be used to obtain taxonomic and functional distributions (8) and allows to directly feed metabolic pathway identification (9) that in turn can be converted into stoichiometric models (10) for simulating the behaviour of single organisms or the relationships within a community, with the potential of predicting their response to changing environmental conditions.

Table 1: Tools for metagenomic analysis indexed by scope

Scope	Name	Link to program
Recruitment	BWA	bio-bwa.sourceforge.net
	Bowtie	bowtie-bio.sourceforge.net
	FR-HIT	weizhong-lab.ucsd.edu/frhit
Assembly	Meta-Velvet	metavelvet.dna.bio.keio.ac.jp
	META-IDBA	i.cs.hku.hk/~alse/hkubrg/projects/metaidba/
	IDBA-UD	i.cs.hku.hk/~alse/hkubrg/projects/idba.ud/
	Genovo	cs.stanford.edu/group/genovo/
Genes	FragGeneScan	omics.informatics.indiana.edu/FragGeneScan
	MGA	whale.bio.titech.ac.jp/metagene
	Glimmer-MG	www.cbcb.umd.edu/software/glimmer-mg
	GeneMark	exon.gatech.edu/metagenome
Annotation	RPSBlast	www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
	HMMer3	hmmer.janelia.org
	BLAST	blast.ncbi.nlm.nih.gov
	RAPSearch2	omics.informatics.indiana.edu/mg/RAPSearch2
	RAST	rast.nmpdr.org/
Taxonomy	RDPclassifier	rdp.cme.msu.edu
	NBC	nbc.ece.drexel.edu
	CARMA3	webcarma.cebitec.uni-bielefeld.de
	MEGAN	ab.inf.uni-tuebingen.de/software/megan
	SOrt-ITEMS	metagenomics.atc.tcs.com/binning/SOrt-ITEMS
Servers	MG-RAST	metagenomics.anl.gov
	IMG/M	img.jgi.doe.gov/
	EBI metagenomics	https://www.ebi.ac.uk/metagenomics/
Models	PathwayTools	bioinformatics.ai.sri.com/ptools/
	Model SEED	seed-viewer.theseed.org/seedviewer.cgi?page=ModelView
Analysis	GSEA	www.broadinstitute.org/gsea/
	ShotgunFunctionalizeR	http://shotgun.math.chalmers.se/
	MetaPath	www.cbcb.umd.edu/~boliu/metapath/
	STAMP	http://kiwi.cs.dal.ca/Software/STAMP
	HUMAnN	uttenhower.sph.harvard.edu/humann

taxonomic information useful as evidence for presence of a set of cellular functions and biochemical pathways. We will then review the methods transforming NGS-derived short sequence reads into taxonomic and functional entities, which in turn can be framed into the context of biological pathways.

BIOLOGICAL PATHWAYS: THE BIOINFORMATICS PERSPECTIVE

A biological pathway is classically defined as the series of molecular interactions that leads to a certain product or cellular function. The two-dimensional graphical display of a pathway aims to capture the interdependencies between elements that concur to a biological function resulting from the sequential interaction of the elements. A biological pathway is the result of a manual curation made by experts in different fields aimed at building networks of genes that have experimentally proven relationships (e.g. substrate–product link, physical association, post-translational modification) and cooperate to a common biological goal. Efforts to create centralized repositories of pathways such as Kyoto encyclopedia of genes and genomes (KEGG) [3] or MetaCyc [4] are struggling with different data models to make pathways homogeneous in terms of representation and coding (see www.pathguide.org for a survey of the wealth of pathway repositories so far available).

From a computational point of view, coding (the framework of rules that allow a pathway to be described textually) is definitely far more important than representation. In fact, a correct parsing of the data that catches information about the elements and their relationships is fundamental to take full advantage of the efforts made by experts in building such sets. A number of alternative schemes, mostly implemented in specialized XML (eXtensible Markup Language), have been proposed to code increased levels of complexity (SBML [5], SBGN [6], BioPax [7], CellML [8], KGML [3], BCML [9]). Pathway repositories also put pathways in the context of

genomic research and taxonomy, trying to predefine the pathways available in an organism based on the genes detected in its genome. Thanks to these repositories, gene expression (transcriptomics), protein expression and modifications (proteomics), primary and secondary metabolites production (metabolomics) as well as their control systems can be framed in the context of flux analysis, reconstructing pathways and modelling their behaviour in a mechanistic and mathematically appropriate framework [10]. Describing biological processes as a function of the connectivity between the elements is even more intriguing in the microbial metagenomics field, where the classic concept of pathway cannot be confined into one organism but has to be reinterpreted in terms of flux of information across different species. The simultaneous analysis of complex microbial communities requires defining inter-organismic 'meta-pathways', as constructed by combining multiple pathway parts from multiple organisms, to highlight the flux of interactions between them and identify the metabolic functions that make a complex microbial community. A striking example of such inter-organismic pathways can be found in a recent analysis of global gene expression of the bacteriocyte *Buchnera aphidicola*, supporting a genome-wide coordination of host gene expression with bacterial metabolic pathways [11]. *Buchnera* produces essential amino acids, such as methionine, that are deficient in the aphid's diet with the help of complementary aphid-encoded enzymes (Figure 2) [12]. Pathway-level integration of the different capabilities of different species reveals how beneficial associations can arise *de novo* from organisms that are not co-evolved and later become stabilized through natural selection acting within each species.

The one above is a clear example of how metagenomics can improve and complement existing views on pathway evolution, yet more frequently metagenomics is used to investigate the taxonomic composition of the environment, extract dominant species and extract their pathways, if available.

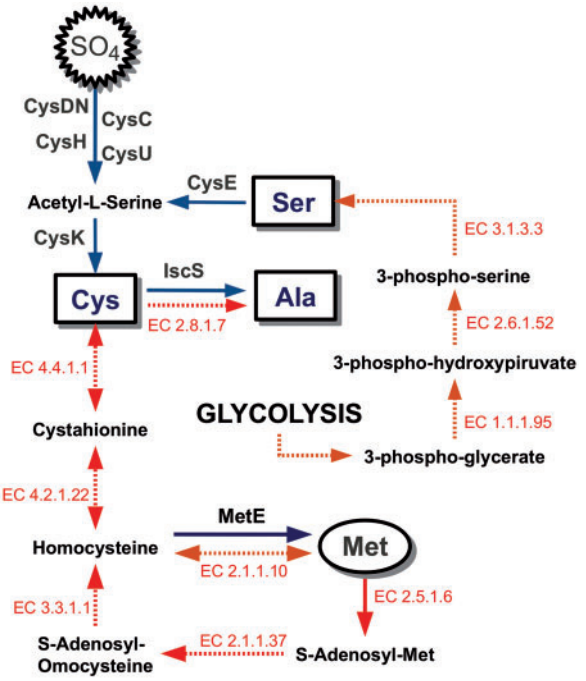

Figure 2: Example of a 'meta-pathway': amino acid biosynthesis in the *Acyrthosphion pisum*/*Buchnera aphidicola* symbiosis. Amino acids in squared boxes are non-essential, methionine in round box is essential. Solid lines and gene names are for *Buchnera*, dashed lines and EC codes are for *Acyrthosphion*. The non-essential amino acid cysteine (Cys) is synthesized by *Buchnera aphidicola* from phloem sap provisioned sulphate and *A. pisum* synthesized serine (non-essential). Adapted from [11].

TAXONOMY AS THE BASIC LAYER OF INFORMATION FOR METAGENOMIC ASSEMBLY

The knowledge of the taxonomic distribution of individuals within a metagenomic sample has a deep impact on the functional assignment of genes. A high biodiversity negatively correlates with the functional assignment, possibly due to the presence of unknown organisms that potentially encode elusive functions. Microbial community analysis obtained with rRNA genes or other markers favours the development of rapid taxonomic classifiers (RDPclassifier [13] and NBC [14], Figure 1, step 1) based on naïve Bayesian statistics and accurate taxonomically organized reference databases (RDP [15], SILVA [16], GreenGenes [17] for bacteria and archaea and AFTOL [18] for Fungi). At an early stage, these databases revolutionized our understanding of how life is organized at the kingdom level on earth and have since provided a powerful research and reference tool for microbial ecologists, microbial taxonomists and applied microbiologists alike. The Roche FLX Genome Sequencer is particularly suited for 16S rRNA-based surveys since it can produce about 1 million high-quality reads of 400–700 bp

with a well-defined base-call error model that facilitates data pre-processing. Samples can be multiplexed, thanks to barcodes applied during the amplification step, allowing accurate high-throughput microbial identification to the species level within metagenomic samples. It should be emphasized that the debated definition of bacterial species, strongly hampered by the extensive exchange of genetic material, could receive a positive improvement by metagenomics [19]. Another widely used approach to investigate community richness and diversity variations, and to complement the taxonomic characterization, is to cluster the sequences according to their distance. Distances can be obtained with or without a multiple sequence alignment [17, 20, 21] allowing to cope with never observed or classified species that are frequent in metagenomics.

The information on the taxonomic composition of a sample can be used to infer the metabolic networks and biochemical pathways present in that sample. Knowing 'who is there' is not necessarily enough to understand natural microbial communities but may serve as a proxy for metabolic reconstructions, since likely the most abundant species are dominant and, if their genomes are known in sufficient details, their metabolic pathways can be used for simulations (see below) [10].

GENOME ASSEMBLING FROM HIGHLY UNEVEN, LOWLY COVERING READS SET

Assembly, i.e. the process of juxtaposing short, overlapping fragments and creating longer sequences, ideally spanning whole chromosomes, is a very complex task in environmental metagenomics. A *de novo* assembly is sometimes possible if the estimated diversity is very low and composed of quite different species. In such cases, the coverage of the sequencing is sufficient to confidently combine reads into longer contigs, frequently after a pre-clustering based on, e.g. base frequency distribution into appropriate independent assembly lots. If a taxonomic survey is available, reference genomes of closely related bacterial strains can be used to guide the assembly process. Algorithms developed for single genomes assembly from short reads are still commonly used to assemble such metagenomic data sets and rely on graph reduction algorithms based on three alternative strategies, namely, Overlap Layout Consensus, De Bruijn Graph (DBG) and the greedy graph. Details of these programs are well and extensively reviewed by Miller *et al*. [22] and an in-depth description is far beyond the scopes of this review. Usually all the approaches produce shorter contigs (both in terms of N50 and maximum contig length) if compared with single genome assembly, mainly because of the very low coverage of sequencing due to the inherent complexity of the sample. Another layer of complexity in metagenomic assembly is due to high frequency of polymorphisms and genome variations that, along with low complexity regions, lead to mis-assembly and chimeric contigs formation and is furthermore hampered by the presence of viruses and inserted phages [23–25]. Recently, novel assemblers specific for metagenomics (i.e. Genovo [26]) started to be developed (Figure 1, step 2). The well-known and widely used DBG-driven assembler Velvet has been adapted to cope with multiple genomes, leading to the release of MetaVelvet [27]. This update introduced the possibility of isolating sub-graphs according to a k-mer coverage histogram, that in metagenomes should present a multi-modal distribution (indicating the presence of different organisms, rather than a unimodal distribution typical of single genomes) and to build scaffolds based on every decomposed de Bruijn sub-graph containing reads from related genomes. A similar task is implemented in Meta-IDBA [28] that partitions the graph into components based on the topological structure of the graph and takes advantage of the iterative approach of the IDBA method that, differently from all other De Brijin graph assemblers, does not rely on a specific and possibly inappropriate k-mer size but builds graphs with a range of them, with benefits for the overall assembly process [29]. A very recent additional improvement to the IDBA system for metagenomes is IDBA-UD that addresses the problem of uneven distribution of the reads by iteratively removing short inaccurate contigs and performing local assembly to fill gaps. This allows larger k-mers to be evaluated, obtaining longer contigs with less gaps in both low-depth and high-depth regions [30]. A completely different point of view has recently been proposed by MetaORFA [31] that introduced the assembly of proteins instead of genes form open reading frames (ORFs) predicted in DNA fragments, using an Euler assembler modified to cope with an amino acid alphabet.

Despite the progress, genome assembly is still an error-prone process and the final result depends mainly on the genome structure and the complexity

of the metagenomic data set. Importantly, the benchmark of metagenomic assembly programs is frequently based on simulated data sets from fully sequenced genomes (e.g. produced with the valuable MetaSim software [32] that can produce simulated reads according to a number of parameters such as taxonomy distribution and technique-based error models), providing a ground-true basis for comparison, but not necessarily representing real data sets, which can contain extremely uneven reads, in particular for unknown species.

FINDING GENES IN METAGENOMES

In metagenomics, there are two contrasting forces in action: the greediness of the functional assignment and the necessity of quantification, i.e. giving counts to the assigned functions. We will address these two topics as a unique ensemble here, to help catching the balance that has to be kept in drawing robust conclusions on metagenomics results.

Gene finding on genomic sequences is a fundamental step which allows the annotation and characterization of the functional potential of the prokaryotic community under investigation. In metagenomics sequencing projects, particularly for complex communities, gene calling is hampered by the fragmentation of the assembly that affects the genome of low abundant species producing also unassembled singletons. Moreover, reads produced by NGS contain errors according to the particular technology used to sequence the genomes: this can lead to frame shifts and make gene prediction more difficult. Classical tools for gene finding on genomes (e.g. Glimmer [33]) efficiently base their predictions on hidden Markov models (HMMs), trained on the gene structure of known similar organisms or on generalized prokaryotic or eukaryotic genes. As metagenomic-derived genes originate from a mixture of different organisms, this approach cannot be used directly, at least not with the confidence used in single genomes. To overcome this problem, gene predictors based on more complex models have been developed that try to minimize the limitations imposed by the lack of predetermined models and incorporate codon bias and start/stop codon patterns of known genes of available whole genomes (Figure 1, step 3). MetaGene Annotator [34] integrates statistical models of bacterial, archeal and prophage genes; uses di-codon-based self-training models selected from input sequences (based on GC content) and incorporates species-specific patterns of ribosome binding sites, allowing increased confidence in predicting translation starts. FragGeneScan [35] instead incorporates in the same HMM codon usage bias, sequencing error models and start/stop codon patterns and allows the recovery of genes directly from short fragments (reads) since it does not need evident start or stop codons. Recently, the well-known Glimmer gene finder has been updated into Glimmer-MG [36] that also integrates phylogenetic classifications and sequence clustering (pre-grouping together those genes that likely originate from the same organism) to further improve gene prediction. Finally, the another widely used HMM-driven gene finder GeneMark has been adapted to take advantage of direct polynomial and logistic approximations of oligonucleotide frequencies from short metagenomics reads to heuristically drive the model parametrization and obtain better gene predictions [37]. To the best of our knowledge, no recent independent comparison has been performed on specificity and sensitivity of metagenomics gene finders, but it has been reported [35] that FragGeneScan has the greatest accuracy with Illumina-sized read length (~100 bp), while longer sequences obtained, e.g. from pyrosequencing or even from assemblers, are predicted with high accuracy by most gene finders. The constraints imposed by the format of this review do not allow performing an extensive comparison of the different gene finding algorithms, yet the importance of this subject indicates that such an effort should indeed be undertaken.

It remains a hard task to distinguish between true ORFs and false ones. An approach used to solve this limitation is progressive clustering of ORFs with calculation of Ka/Ks (synonymous and non-synonymous substitution rates), assessing the selective pressure acting on the ORF and, if absent, to score it as unlikely (a possible false positive) [38]. A limitation of the Ka/Ks approach for metagenomic data sets is the lack of reference genomes for several of the sequences and the consequent uncertainty in mapping the ORFs to chromosomes and discriminating homologs from paralogs. A variety of web-oriented computational resources are also available such as RAMMCAP [39], which makes use of CD-HIT [40] as a fast sequence clustering method to group similar ORFs and reduce the amount of data while increasing reliability.

FUNCTIONAL ANALYSIS ON GENES OR READS

A possible approach for function finding and quantification in metagenomics is to treat reads such as transcriptomics data and try to simply map them to reference genomes (e.g. versus NCBI nt or MetaHIT unique CDS), counting the number of matches and scoring the functions accordingly (Figure 1, step 4). A number of programs have been developed to rapidly and efficiently accomplish this task using reads recruiters, i.e. very fast pair wise aligners (Figure 1, step 5) such as BWA [41], Bowtie [42] (both built upon the Burrows–Wheeler Indexing system) or FR-HIT [43] (that builds a k-mer hash table for the reference sequences and then performs seeding, filtering and banded alignment to identify the alignments to reference sequences that meet user-defined cutoffs) or more classic, parallel versions of BLAST [44]. Due to the enormous amount of information to be processed, speed is an important aspect of recruiters even if false-positive results are frequent due to the highly heuristic procedures. A good balance seems to be present in FR-HIT, which shows accuracy similar to the slower BLAST but consistently higher than the faster BWA and Bowtie, even if independent studies have not been drawn so far. Such raw quantification is usually followed by normalization for reference coding sequence length [45]. This approach risks to be hampered by sequence conservation due to functional homology in different organisms. In fact, a read that maps into a highly conserved region of a gene (e.g. in a structurally or functionally conserved region of the coded protein, that tend to be highly maintained in evolution) will probably be assigned to different targets with a similar score.

A possible solution is to move from CDS-based references to profile-based protein references. A small number of existing databases collects multiple sequence alignment of protein sequences that share a proved (experimentally) or predicted (from sequence similarity) function. Probably the most important example of such databanks is NCBI Conserved Domain Database [46] that incorporates proteins from several sources such as Pfam/TIGRFam, COG, Prk, Cazy and many others and FIGfams, which rely on the SEED classification and grouping of elements in sequenced genomes. Fast search engines have been developed to scan protein sequences against HMMs or profiles generated from multiple sequence alignments (e.g. HMMer3 [47] or RPS-Blast [42]) and are currently in use for functional assignment directly on reads or genes from metagenomics assemblers. Another possibility is to scan protein databases such as NCBI nr (non-redundant) with fast, specifically designed protein search tools such as RAPsearch2 [48] that use reduced amino acid alphabets (by combining residues with similar characteristics in a common symbol) to reduce the overall complexity of the search while maintaining (an sometimes improving) the recovery of low-similarity proteins.

PATHWAY ANALYSIS AS A TOOL TO COMPARE METAGENOMICS POPULATIONS

The development of methods for pathway-based analysis has spearheaded the application of bioinformatics in the functional genomics field, becoming the tool of choice for 'guilt by association' statistical analyses. Such methods allow the linking of a phenotype to a cascade of events stemming from a series of connected sets of genes or proteins. Microarray-based gene expression studies have used and abused pathway analysis often calling a pathway what is not a pathway, but rather a gene set and determined a functional enrichment statistics or score based on a pool of genes in the sample (e.g. Fisher exact test, gene set enrichment analysis [49]). The result of enrichment methods is a list of pathways that appear to be significantly over-represented. All these count-based approaches do not take into account neither the relative order of the connected genes nor the topology of the underlying graph: they can help in interpreting the results, but cannot address the effective behaviour of the pathway. More complex methods have therefore been developed that address the topology and the connectivity of the elements in the pathway, as well as the flux of changes following a challenge. Methods such as Impact Analysis can effectively determine the activation/inactivation status of a pathway by appropriately weighting the perturbation that an over- or down-representation of an element of the pathway has on the whole pathway. This represents the state of the art of pathway analysis in gene expression studies of mammalian cells or model organisms [50].

Several drawbacks limit the application of statistical methods measuring pathway enrichment to metagenomics data. The main limitation is that most of the methods applied to metagenomic data

sets to score differences in pathway abundance have been developed for microarrays. Specifically tailored tools should take greater care in evaluating missing values that are frequent in metagenomic data sets. So far metagenomics studies have been mainly measuring DNA sequence abundance, only a few studies, such as the *Buchnera* one [12], measure gene expression, consequently several of the statistical assumptions of the methods analysing gene expression profiles in pure cultures are not applicable. Most importantly, NGS studies investigating the role of symbiotic microbial populations in fundamental biochemical functions hardly ever discover in any given sample all the genes making up a pathway. The difficulty in measuring all the components of a given pathway results in following simple rules of thumb based on the minimal number of elements sufficient for a pathway to be considered as present (Figure 1, step 6). A frequent assumption is that if an element of a pathway is present in a gene set, this pathway can be considered as present and scored accordingly (the so-called naïve approach). It can be argued that if a single gene contributes to two pathways automatically makes the score of both pathways increase: this behaviour raises doubts about their real copy number. The opposite attitude is instead to consider a pathway as present only if all its elements are found in a sample. This approach is instead too conservative and even less intuitive: it is clear that pathways are 'ideal' groups of genes whose number have been arbitrarily set from a functional point of view and all the functional elements actually coexist, possibly at the same time.

A possible solution to the problem of missing values in metagenomic pathway analysis is reducing their size into smaller modules with more detailed, specific features. The KEGG database maintains ancillary, modular subsystems that compose pathways and can be instanced separately. Flux analysis frequently searches for reduced versions of pathways with the aim of reducing the total computational burden of the simulation (examples are the aforementioned Elementary Modes, Paths or Patterns). Biology also offers examples of modularity in bacterial operons, bacterial genes of a metabolic pathway organized in the same messenger RNAs and therefore co-regulated. Methods and tools for operon prediction and bacterial pathway reconstruction, resolving function distribution within single circular bacterial genomes, have been recently discussed [51]. Their mapping in metagenomics seems easier due to their relative smaller size, but they are constrained to respect the assumption of co-linearity, and therefore have to rely on the performance of metagenomic assemblers, that though promising often implement conservative assumption and potentially fail calling all the genes of an operon as present. The task of operon and pathway reconstruction is made even more complicated since genomes are known to have gaps in the commonly used pathways [52] and even if in metagenomics many bacterial species are present, the coverage on the genome is usually inversely proportional to the biodiversity of the sample, the higher the biodiversity the lower the coverage. Some intermediate solutions have been proposed to adjust pathway abundance and to avoid overestimation. The PathoLogic module of BioCyc PathwayTools offers a machine learning approach to pathway assignment based on a large number of features (pathway composition and connectivity, genomic context and pathway variants, plus manually imposed constraints) learned from some accurate complete genomes [53, 54], leading to accurate predictions on single genomes (>91%). MinPath [55] introduced a parsimony method solved with integer programming to filter out spurious assignments and basically to find the minimal set of pathways that can be explained with the supplied gene functions and abundances. This approach lead to a conservative yet accurate identification of pathways in single genomes when compared with KEGG identifications (that are known to be inflated) and does not rely on training, so it appear more suitable for metagenomic data sets. In fact, the current number of complete bacterial genomes is still too limited with respect to the worldwide microbial diversity and the dependence of PathoLogic on completeness of the genome as a source of information (e.g. the genomic context of the genes or the taxonomic similarity) partially limits its application in metagenomic data sets, which mainly contain sequences from hypothetical non-cultivable taxonomical entities. Metagenomic-specific methods usually assess their accuracy by constructing synthetic read/gene sets from existing complete genomes and simulate communities at different complexities (e.g. using the MetaSim software [32]). This kind of benchmark sets represents the golden standard for metagenomic tailored programs.

Recently, a statistical framework was proposed for modelling gene family abundance [56] and two models for pathway analysis were elaborated taking

into account pathway size, gene length and gene overlap, using genes known to be present only once per genome [57, 58] for normalization. Unfortunately, to the best of our knowledge these models have not been distributed as usable software and therefore they represent interesting yet theoretical methods. Finally, when calculating statistical enrichment, the study design requires determining a ratio between conditions. Metagenomics studies have barely defined the rules to normalize across samples and within the same sample (Figure 1, step 7). A confident quantification of gene abundance is fundamental when comparing the results from different metagenomic samples, since incorrect mapping is the major source of overestimation and bias. Normalization of metagenomic data requires to account for the estimate of average genome sizes, relieving comparative biases introduced by differences in community structure, number of sequencing reads and sequencing read lengths between different metagenomes, as well as from sub-sampling [59, 60]. When discussing the statistical tools to calculate enrichment in functions, so far principal component analysis and non-linear multidimensional scaling are used to visualize the data and identify the factors that characterize different data sets. In 2009, the R package ShotgunFunctionalizeR was published, which allows gene- and pathway-centric analyses based on statistical analysis such as binomial and hypergeometric tests and generalized linear models with a Poisson canonical logarithmic link [61]. In 2010, the STAMP project was presented [62] showing a graphical interface for metagenomic analyses and, most notably, an open community was created that promotes 'best practices' in choosing appropriate statistical techniques and reporting results in metagenomics. In 2011, the LEfSe statistical procedure was introduced, implementing multivariate techniques to robustly identify features that are statistically different among biological classes and then performing additional non-parametric pairwise tests to assess whether these differences are consistent with biological test cases, also providing size effects and dimension reduction to the results sets [63]. A probably more pathway-oriented analysis was proposed in 2010 with MetaPath that included the network structure into statistics. Starting from the KEGG global metabolic pathway (actually from the network of KEGG reactions), MetaPath uses Metastats [64] to identify sub-networks that differentiate two meta-samples and it provides statistics for pathway abundance and topology.

Very recently a new promising methodology to reconstruct the functional potential of microbial communities from metagenomic sequences was proposed with the name HUMAnN (HMP Unified Metabolic Analysis Network) [65]. The authors propose a combination of several of the above-mentioned steps but they add several improvements, among which: (i) filtering steps to ensure that unlikely pathways are removed and that the abundance of consistent pathways is robustly evaluated, (ii) a normalization step based on taxonomic profiles from BLAST hits and (iii) a combination of pathway abundance and coverage (i.e. the proportion of genes in the pathway actually found in the sample) to appropriately interpret the results.

METABOLIC NETWORKS RECONSTRUCTION AND SIMULATIONS

The increasing availability of complete, annotated genomes, allows building genome-scale metabolic models (GMMs) directly from abundance analyses at the species or genus level.

The methods that will be described in this paragraph rely heavily on the accuracy of the taxonomic annotation. Concerns have been raised regarding the use of 454 pyrosequencing for determination of taxonomic abundance, due to the biasing effect of using a reference database or to the amplification step required in DNA preparation for pyrosequencing. These limitations can be minimized by using multiple databases and by confirming major outcomes using PCR-independent methods, including NGS techniques that do not require amplification steps. Tools have therefore been developed that use protein or nucleotide databases to extract taxonomic data from short metagenomic sequence fragments (with or without prior assembling, Figure 1, step 8). Examples of such tools are MEGAN [66], CARMA [67], Sort-ITEMS [68] or MetaPhyler [69] that use similarity searches, the faster PhyloPythia [70] or TETRA [71] that use taxonomy-fitted nucleotide or codon usage composition analysis or SPHINX that uses a combination of the two [72].

A number of models have already been published [73] by integrating known interactions (e.g. reagent/product at the enzyme level) from pathway

repositories such as KEGG [3] or MetaCyc [4] with stable annotations, e.g. in UniProt [74] or Brenda [75] databases (Figure 1, step 9). This knowledge-based integration can be converted into a mathematical model that can be analysed through constraint-based approaches and linear programming methods with one or more objective functions (e.g. consumption/production of a metabolite or, more frequently growth rate upon medium change). In 2010, a semi-automatic model generation system has been developed [76] based on the RAST annotation system upon the SEED framework (relying on FIGfam database) that automatically maps genes from full genomes into metabolic maps connected with the KEGG database (Figure 1, step 10). This system frames each gene in an appropriate metabolic context and incorporates, beside pathway topologies (with enzymes as nodes and reactions as directed edges), a wide information on intervening compounds and the notion of essentialness that helps in filling gaps, genes that have not been detected but that must be present as a function for a pathway to be rigorously defined. Once the model has been generated, a battery of techniques can be used to reduce its complexity and impose biological, spatial or thermodynamic constraints to find optimal metabolic states via flux balance analysis (FBA). Toolboxes that accomplish most of these tasks are COBRA, excellently reviewed in [77] and PathwayTools, that use the BioCyc models and take full advantage of the MetaCyc pathway database [4]. Searching for elementary flux modes (i.e. the minimal number of enzymes that works at steady state with all irreversible reactions pointing to a given end) is a complex task since their number grows exponentially with GMM size. This problem can now be addressed, thanks to linear programming [78], but alternatives that reduce the complexity exist such as elementary flux patters [79] or flux path [80].

All the above mentioned methods and theories have mostly been developed and applied to single cells, but several attempts have been proposed in which stoichiometric GMMs from different organisms are mixed following the already established rules for modelling sub-cellular compartments [81]. Examples are the pioneer works on mutualistic relationships between *Desulfovibrio vulgaris* and *Metanococcus maripaludis* [82] and on synthropies between oxygenic phototrophs, filamentous anoxygenic phototrophs and sulphate reducing bacteria of the Yellowstone National Park (USA) [83], as well as the demonstration that modelling can be used to identify media that stimulate symbiotic relationships [84]. Other important examples can be found in the excellent review of Klitgord and Segrè [85]. Importantly, increasing interest in microbial community metabolic simulations is evident, as shown by the recent development of OptCom, a microbial community-addressed FBA framework [86].

Metagenomics offers the possibility of complementing existing whole genome sequences by determining, after appropriate processing and through a number of specifically designed bioinformatic tools, the true presence/absence of nodes in the pathways and eventually to add nodes previously unknown, therefore greatly improving the precision of reconstructed models.

Application of more than one taxonomy inference method is likely to improve the reliability of GMMS and other taxonomy-based network reconstruction tools. On the other hand, Illumina-based metagenomics provides the investigator with an enormous amount of sequence information on metabolic networks, allowing moving from inference to measurement of the abundance of genes and transcripts. Hereinafter, we will address key steps and pitfalls of the process for extracting functional information from Illumina metagenomics reads.

INTERNATIONAL EFFORTS AND RESOURCES FOR METAGENOMICS PATHWAYS

As happened in the past, with DNA sequences and microarray data, repositories specialized in classifying and organizing metagenomics data have arisen, thanks to the efforts and funding of international consortia [87]. The ability to share metagenomics data requires the definition of the minimal set of information that has to be made available to the community to allow comparing data sets from different laboratories [88, 89]. The definition of common standards is a prerequisite for the development of new analysis methods to be tested on a sufficiently large and robust benchmark. The process of defining such standards is of paramount importance to achieve the final goal of improving the structure and dynamics of microbial communities and their relationships with ecosystems, natural or artificial habitats and, importantly, human biology and pathobiology.

The amount of metagenomic information is exponentially increasing, the first EU-funded MetaHIT consortium produced Illumina sequences of faecal samples of 124 European individuals, including healthy, overweight and obese adults as well as patients with inflammatory bowel disease [90]. When extended to Japanese and American populations, MetaHIT also established that worldwide population could be classified into three distinct enterotypes [91]. The NIH-funded Human Microbiome Project is also curating and indexing another fundamental resource for metagenomics, i.e. a catalogue of reference genomes hosted in the Genome Online Database framework [92], that similarly to other tools such as MeganDB, MG-RAST [93], IMG/M [94] or Camera offers web services dealing with data pre-processing, assembly, gene finding, functional assignment and, in some cases, pathway reconstruction. The information deposited in these resources promises to be a goldmine for pathway and network inference, reconstructing the super-meta-pathway subtending the interaction between mammals and their microbiomes. A glimpse of the metabolic pathway complexity contained in metagenomics data sets appeared since the work of Gill *et al*. [95]: the human genome lacks most of the enzymes required for degradation of plant polysaccharides and they are supplied by the human gut microbiome that can metabolize cellulose, starch and unusual sugars such as arabinose, mannose and xylose, thanks to at least 81 different glycoside hydrolase families.

Zhu *et al*. [96] undertook a large-scale analysis of 16S rRNA gene sequences to profile the microbiota inhabiting the digestive system of giant pandas using a metagenomic approach. They performed predicted gene functional classification by querying protein sequences of the genes against the eggNOG database (an integration of the COG and KOG databases) and the KEGG database using BLASTP, finding the presence of putative cellulose-metabolizing symbionts in this little-studied microbial environment, explaining how giant pandas are able to partially digest bamboo fibre despite a genome lacking enzymes that can degrade cellulose. Recently, Segata *et al*. [97] introduced an innovative analysis of the HMP metagenomic shotgun sequencing of a subset of the available body habitats (adult digestive tract). The study design offered an additional feature, the measure of the relative abundances of bacterial organisms based on 16S rRNA genes. The authors examined the abundances of microbial metabolic pathways including the relative abundances of individual enzyme families Kyoto encyclopaedia of genes and genomes (KEGG), Orthologous groups and of complete metabolic modules, identifying a core set of metabolic pathways present across these diverse digestive tract habitats. The application [98] of an ensemble method based on multiple similarity measures in combination with generalized boosted linear models to taxonomic marker (16S rRNA gene) profiles of the HMP cohort resulted in a global network of 3005 significant co-occurrence and co-exclusion relationships between 197 clades occurring throughout the human microbiome. This network revealed strong niche specialization, with most microbial associations occurring within body sites and a number of accompanying inter-body site relationships. The co-occurrence of microbial species in similar abundance could be seen as an indication of their being part of an integrated network, providing a set of mutually complementary functions integrated in a multi-organismal pathway. The size of this super-network can be estimated integrating the HMP and the MetaHIT studies, indicating that gene content in gut microbiota is at least 150-fold higher than human genome and identifies >19 000 different functions, among which at least 5000 never seen before, at least 6000 shared by all individuals (the so-called 'minimal metagenome') and at least 1200 required for any bacterium to strive in the human gut (the 'minimal gut microbiome'). Finally, it is becoming increasingly clear that this network varies significantly with geography and diet and we are just starting to appreciate its complexity. Yatsunenko *et al*. [99] analysed the gut microbiome from healthy children and adults from the Amazonas of Venezuela, rural Malawi and US metropolitan areas. KEGG ECs data analysis found that largest differences were determined by Random Forests and ShotgunFunctionalizeR analyses, finding pronounced differences in bacterial assemblages and in the functional profiles in the three study populations, with distinctive features evident in early infancy as well as adulthood.

CONCLUSIONS AND PERSPECTIVES

The new landscape we face requires a profound rethinking of our definition of pathway,

as well as the development of a next generation of pathway data models for the metagenomics field.

Biochemical and regulatory pathways have so far been thought and modelled within one cell type, one organism, one species. Recently, cell type-specific pathway databases and data models were developed to dissect the contribution of different cell types to immune function [100]. With the advent of whole microbiome sequencing studies this cell type-specific pathway annotation paradigm will be generalized to annotate species-specific or strain-specific reactions as part of integrated cross-species 'super-meta-pathways'. In this novel perspective, the network stemming from the interaction of a community of cells includes all the functions of the cell type that make the system, irrespectively of the species contributing a function or a set of functions. This novel pathway annotation will reconstruct and model biochemical and regulatory pathways as the result of integration of function in complex symbiotic interactions, indicating exactly which metabolite is the end product in one cell type and how this metabolite enters a cell type or a microorganism. The elements of the network will have to be functionally connected by means of tools such as those integrating gene expression with metabolite networks [101], with an approach conceptually similar to that developed by De Filippo *et al*. [102] correlating the presence of short chain fatty acids (SCFAs) in a sample with the species carrying the genes for the pathways involved in SCFAs production from their precursors. Currently, the principal bottleneck for the progress of pathway analysis of metagenomics data remains assignment of function and assembly of operons and metagenomes. Despite the fast paced advancement of tools for statistical pathway analysis of metatrascriptomes or metagenomes, the lack of methods to fill gaps, and of a proper pathway data model integrating reactions present in taxonomically different organisms, makes difficult to determine the exact topology of the pathway, thus limiting the application of the most advanced topology-based methods to metagenomics. Drastic improvement will be in the future driven by the appearance of technologies increasing the length of the sequencing reads, while maintaining or increasing the throughput and their application to the analysis of gene expression in mixed samples.

Key Points

- Next-generation sequencing is cradle and stage of the metagenomics revolution.
- Pathways are uniform biological framework that can be used for functional modelling of microbial metabolisms.
- Taxonomy is the basic layer of information for metagenomics analyses.
- Metagenomics data can supplement, complement and refine pathway reconstruction.
- Metagenomics assembly is one of the most complex tasks that bioinformatics is facing today.
- Finding genes in metagenomes is of paramount importance for improving current methods for pathway analysis.
- Functional assignment of metagenomics data is a complex and error-prone task that struggle against incompleteness of both raw data and current available information on genes and proteins.
- Genome-scale metabolic models and taxonomy can be combined into mathematical modelling for single cell and community-based analyses.
- Inferred pathways can be used as a tool to compare metagenomics populations and to rationally build models of microbial communities by considering them as communicating elements of a single, self-comprehensive map of biochemical reactions.

FUNDING

This work was supported by grants from the EU FP7 Integrative project SYBARIS [242220].

References

1. Allen EE, Banfield JF. Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* 2005;**3**:489–98.
2. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;**11**(1):31–46.
3. Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;**40**:D109–14.
4. Caspi R, Altman T, Dale JM, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2010; **38**:D473–9.
5. Hucka M, Finney A, Sauro HM, *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**(4):524–31.
6. Le Novère N, Hucka M, Mi H, *et al*. The systems biology graphical notation. *Nat Biotechnol* 2009;**27**(8):735–41.
7. Demir E, Cary MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**(9): 935–42.
8. Miller AK, Marsh J, Reeve A, *et al*. An overview of the CellML API and its implementation. *BMC Bioinformatics* 2010;**11**:178.
9. Beltrame L, Calura E, Popovici RR, *et al*. The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics* 2011;**27**(15):2127–33.

10. Oberhardt MA, Puchałka J, Martins dos Santos VA, Papin JA. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS Comput Biol* 2011;**7**(3):e1001116.
11. Shigenobu S, Wilson AC. Genomic revelations of a mutualism: the pea aphid and its obligate bacterial symbiont. *Cell Mol Life Sci* 2011;**68**(8):1297–309.
12. Hansen AK, Moran NA. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proc Natl Acad Sci USA* 2011;**108**(7):2849–54.
13. Wang Q, Garrity GM, Tiedje JM, *et al*. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;**73**(16):5261–7.
14. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 2011;**27**(1):127–9.
15. Cole JR, Wang Q, Cardenas, *et al*. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 2009;**37**:D141–5.
16. Pruesse E, Quast C, Knittel K, *et al*. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;**35**(21):7188–96.
17. DeSantis TZ, Hugenholtz P, Larsen N, *et al*. Greengenes, a chimera-checked 16 S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006;**72**(7):5069–72.
18. Celio GJ, Padamsee M, Dentinger BT, *et al*. Assembling the Fungal Tree of Life: constructing the structural and biochemical database. *Mycologia* 2006;**98**(6):850–9.
19. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. *Environ Microbiol* 2012;**14**(2):347–55.
20. Schloss PD, Westcott SL, Ryabin T, *et al*. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009;**75**(23):7537–41.
21. Cai Y, Sun Y. ESPRIT-Tree: hierarchical clustering analysis of millions of 16 S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res* 2011;**39**(14):e95.
22. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010;**95**(6):315–27.
23. Raes J, Foerstner KU, Bork P. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* 2007;**10**(5):490–8.
24. Charuvaka A, Rangwala H. Evaluation of short read metagenomic assembly. *BMC Genomics* 2011;**12(Suppl. 2)**:S8.
25. Pignatelli M, Moya A. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One* 2011;**6**(5):e19984.
26. Laserson J, Jojic V, Koller D. Genovo: de novo assembly for metagenomes. *J Comput Biol* 2011;**18**(3):429–43.
27. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* 2012;**40**(20):e155.
28. Peng Y, Leung HC, Yiu SM. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 2011;**27**(13):i94–101.
29. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA—A Practical Iterative de Bruijn Graph De Novo Assembler. In: *Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, 2010*. Abstract, pp. 428–440. Springer-Verlag Berlin Heidelderg, Germany.
30. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012;**28**(11):1420–8.
31. Ye Y, Tang H. An ORFome assembly approach to metagenomics sequences analysis. *J Bioinform Comput Biol* 2009;**7**(3):455–71.
32. Richter DC, Ott F, Auch AF, *et al*. MetaSim—a sequencing simulator for genomics and metagenomics. *PLoS ONE* 2008;**3**(10):e3373.
33. Delcher AL, Harmon D, Kasif S, *et al*. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;**27**(23):4636–41.
34. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* 2008;**15**(6):387–96.
35. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 2010;**38**(20):e191.
36. Kelley DR, Liu B, Delcher AL, *et al*. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res* 2012;**40**(1):e9.
37. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 2010;**38**(12):e132.
38. Yooseph S, Li W, Sutton G. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinformatics* 2008;**9**:182.
39. Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 2009;**10**:359.
40. Huang Y, Niu B, Gao Y, *et al*. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**(5):680–2.
41. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.
42. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**(3):R25.
43. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
44. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

45. Kunin V, Copeland A, Lapidus A, *et al*. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* 2008;**72**(4): 557–78.
46. Marchler-Bauer A, Lu S, Anderson JB, *et al*. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 2011;**39**(Database issue):D225–9.
47. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;**39**:W29–37.
48. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 2012;**28**(1):125–6.
49. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.
50. Draghici S, Khatri P, Tarca AL, *et al*. A systems biology approach for pathway level analysis. *Genome Res* 2007; **17**(10):1537–45.
51. Brilli M, Fani R, Lio P. Current trends in the bioinformatic sequence analysis of metabolic pathways in prokariotes. *Brief Bioinformatics* 2007;**9**(1):34–45.
52. Osterman A, Overbeek R. Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol* 2003;**7**(2):238–51.
53. Karp PD, Paley SM, Krummenacker M, *et al*. PathwayTools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinformatics* 2009;**2**(1): 40–79.
54. Dale JM, Popescu L, Karp PD. Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics* 2010; **11**:15.
55. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* 2009;**5**(8):e1000465.
56. Sharon I, Bercovici S, Pinter RY, *et al*. Pathway-based functional analysis of metagenomes. *J Comput Biol* 2011;**18**(3): 495–505.
57. Mollet C, Drancourt M, Raoult D. rpoB sequence analysis as a novel basis for bacterial identification. *Mol Microbiol* 1997;**26**(5):1005–11.
58. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004;**304**(5667):66–74.
59. Frank JA, Sørensen SJ. Quantitative metagenomic analyses based on average genome size normalization. *Appl Environ Microbiol* 2011;**77**(7):2513–21.
60. Cárcer DA, Denman SE, McSweeney C, Morrison M. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. *Appl Environ Microbiol* 2011;**77**(24): 8795–8.
61. Kristiansson E, Hugenholtz P, Dalevi D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 2009;**25**(20):2737–8.
62. Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 2010;**26**(6):715–21.
63. Segata N, Izard J, Waldron L, *et al*. Metagenomic biomarker discovery and explanation. *Genome Biol* 2011;**12**(6):R60.
64. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009;**5**(4): e1000352.
65. Abubucker S, Segata N, Goll J, *et al*. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 2012;**8**(6):e1002358.
66. Huson DH, Mitra S, Ruscheweyh HJ, *et al*. Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 2011;**21**(9):1552–60.
67. Gerlach W, Jünemann S, Tille F, *et al*. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* 2009; **10**:430.
68. Monzoorul Haque M, Ghosh TS, Komanduri D, *et al*. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* 2009;**25**(14):1722–30.
69. Liu B, Gibbons T, Ghodsi M, *et al*. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* 2011;**12(Suppl. 2)**:S4.
70. McHardy AC, Martín HG, Tsirigos A, *et al*. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 2007;**4**(1):63–72.
71. Teeling H, Waldmann J, Lombardot T, *et al*. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 2004;**5**:163.
72. Mohammed MH, Ghosh TS, Singh NK, *et al*. SPHINX—an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics* 2011;**27**(1):22–30.
73. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 2009;**5**:320.
74. UniProt Consortium Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2012;**40**:D71–5.
75. Söhngen C, Chang A, Schomburg D. Development of a classification scheme for disease-related enzyme information. *BMC Bioinformatics* 2011;**12**:329.
76. Henry CS, DeJongh M, Best AA, *et al*. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 2010;**28**(9):977–82.
77. Lewis E, Nagarajan H, Palsson BO. Constraining metabolic genotype-phenotype relationships using a phylogeny of in silico methods. *Nat Rev Microbiol* 2012;**10**(4): 291–305.
78. de Figueiredo LF, Podhorski A, Rubio A, *et al*. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* 2009;**25**(23):3158–65.
79. Kaleta C, de Figueredo LF, Shuster S. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res* 2009;**19**(10):1872–83.
80. Pey J, Prada J, Beasley JE, Planes FJ. Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biol* 2011;**12**:R49.

81. Kiltgord N, Segrè D. The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. *Genome Informatics* 2010;**22**:41–55.
82. Stolyar S, Van Dien S, Hillesland KL, *et al*. Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol* 2007;**3**:92.
83. Taffs R, Aston JE, Brileya K, *et al*. In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC Syst Biol* 2009;**3**:114.
84. Klitgord N, Segrè D. Environments that induce synthetic microbial ecosystems. *PLoS Comput Biol* 2010;**6**(11): e1001002.
85. Klitgord N, Segrè D. Ecosystems biology of microbial metabolism. *Curr Opin Biotechnol* 2011;**22**(4):541–6.
86. Zomorrodi AR, Maranas CD. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput Biol* 2012;**8**(2): e1002363.
87. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, et al. A catalog of reference genomes from the human microbiome. *Science* 2010;**328**(5981):994-9.
88. Sterk P, Hirschman L, Field D, Wooley J. Genomic standards consortium workshop: metagenomics, metadata and metaanalysis (m3). *Pac Symp Biocomput* 2010;481–4.
89. Pan Y, Bodrossy L, Frenzel P, *et al*. Impacts of inter- and intralaboratory variations on the reproducibility of microbial community analyses. *Appl Environ Microbiol* 2010;**76**(22): 7451–8.
90. Qin J, Li R, Raes J, *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;**464**:59–65.
91. Arumugam M, Raes J, Pelletier E, *et al*. Enterotypes of the human gut microbiome. *Nature* 2011;**473**(7346):174–80.
92. Pagani I, Liolios K, Jansson J, *et al*. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2012;**40**:D571–9.
93. Glass EM, Wilkening J, Wilke A, *et al*. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc* 2010;**2010**(1). pdb.prot5368.
94. Markowitz VM, Ivanova NN, Szeto E, *et al*. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* 2008;**36**:D534–8.
95. Gill SR, Pop M, Deboy RT, *et al*. Metagenomic analysis of the human distal gut microbiome. *Science* 2006;**312**(5778): 1355–9.
96. Zhu L, Wua Q, Daia J, *et al*. Evidence of cellulose metabolism by the giant panda gut microbiome. *Proc Natl Acad Sci USA* 2011;**108**(43):17714–9.
97. Segata N, Kinder Haake S, Mannon P, *et al*. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol* 2012;**13**:R42.
98. Faust K, Sathirapongsasuti JF, Izard J, *et al*. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*. 2012;**8**(7). e1002606. Epub 12 July 2012.
99. Yatsunenko T, Rey EF, Manary MJ, *et al*. Human gut microbiome viewed across age and geography. *Nature* 2012;**486**:222–7.
100. Cavalieri D, Rivero D, Beltrame L, *et al*. DC-ATLAS: a systems biology resource to dissect receptor specific signal transduction in dendritic cells. *Immunome Res.* 2010;**6**:10.
101. Faust K, Croes D, van Helden J. Prediction of metabolic pathways from genome-scale metabolic networks. *Biosystems* 2011;**105**(2):109–21.
102. De Filippo C, Cavalieri D, Di Paola M, *et al*. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci USA* 2010;**107**(33):14691–6.