Journal of
Data Mining in Genomics & Proteomics

Bzhalava and Dillner. J Data Mining in Genom Proteomics 2013, 4:3
DOI: 10.4172/2153-0602.1000134

Review Article Open Access

Bioinformatics for Viral Metagenomics

Davit Bzhalava and Joakim Dillner*

Departments of Laboratory Medicine, Medical Epidemiology & Biostatistics, Karolinska Institutet and Department of Pathology, Karolinska Hospital, Stockholm, Sweden

Abstract

Detection of the presence of known and unknown viruses in biospecimens is today routinely performed using viral metagenomics. Because the sequencing speed and cost per base is rapidly declining with new next generation sequencing technologies, such as HiSeq (Illumina), 454 GS FLX (Roche), SOLiD (ABI) and Ion Torrent Proton (Life Technologies), the bioinformatics analysis is today a most important and increasingly demanding part of viral metagenomics analysis. In this review, we highlight some of the major challenges and the most commonly adapted bioinformatics tools for viral metagenomics.

Keywords: Virus; Bioinformatics; Metagenome; High throughput sequencing

Introduction

During past decade we have seen dramatic evolution of next generation sequencing (NGS) instruments like Genome Analyzer/HiSeqSystem (Illumina), 454 GS FLX (Roche), SOLiD (ABI) and Ion Torrent Proton (Life Technologies). A variety of bench top NGS instruments, e.g. the 454 GS Junior (Roche), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies) are now becoming standard equipment in virological laboratories. The performance of various NGS technologies has been reviewed elsewhere [1,2].

NGS technologies can be used to obtain a comprehensive and unbiased sequencing of the DNA present in a sample, without the requirement of any prior PCR or other amplification that requires prior information about sequences that may be present [3]. The complete sequencing of all microbiological sequences that may be present in a sample is termed metagenomics [4]. Viral metagenomics is nowadays routinely used for virus detection and discovery of new viruses [5-13].

As previously pointed out, viral metagenomics has the potential to further our knowledge of the role of viruses in human diseases such as cancer [14]. The last few decades have led to the realization that a considerable proportion of cancers are caused by infections and have also provided epidemiological indications that additional cancer-associated infections may exist [14]. With viral metagenomics, it is possible to perform a large-scale analysis of all infections that are present in cancers and in healthy individuals [14]. Sequencing of cancer specimens with NGS has already been used in the discovery of a new cancer-associated virus, MCV [15]. The Human Microbiome Project (HMP) is one of several international efforts to take advantage of metagenomic analysis and measure microbial diversity in microbiomes from healthy and diseased individuals [16].

Modern NGS technologies are capable of generating billions of bases, at a rapidly decreasing cost per base [1,2]. This increases the demands on the bioinformatics for the analysis of data produced by NGS instruments. In this paper, we review some of the most commonly adapted bioinformatics tools for viral metagenomic analysis, from quality filtering to genome assembly and taxonomic classification.

Bioinformatics Pipeline

The bioinformatics analysis of NGS data for viral metagenomics follows a number of distinct steps, as schematically depicted in Figure 1. This review mostly follows the procedures used in our previous publications [10-13], but with consideration of alternatives and possible improvements.

Quality checking and filtering

The bioinformatics pipelines to analyze next-generation sequencing data usually start by quality checking. The sequences are trimmed according to their Phred quality scores [17]. Phred quality scores are logarithmically related to the base-calling error probabilities. For example, a Phred quality score of 10 corresponds to a base calling accuracy of 90% (10 errors per 100 bp), while quality score of 20 equals to base calling accuracy of 99% (1 error per 1000 bp) [17]. Specific quality filtering conditions can be adapted for different downstream analyses [18].

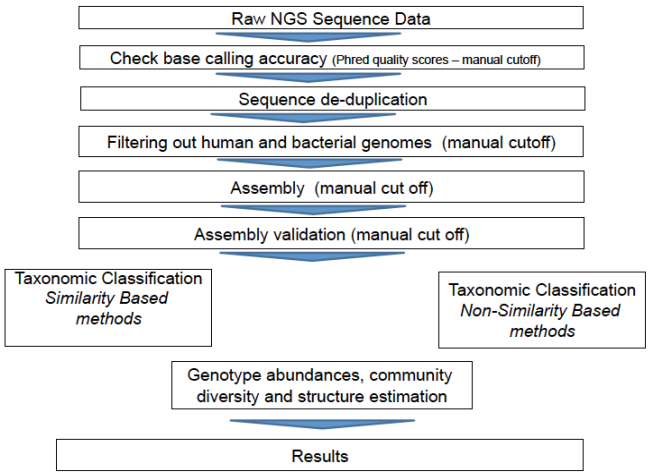

Figure 1: Bioinformatics pipeline for viral metagenomics.

***Corresponding author:** Joakim Dillner, Department of Laboratory Medicine, KarolinskaInstitutet, Huddinge campus F56, Stockholm, Sweden, Tel: +46 76 8871126; Fax: +46 40 337312; E-mail: joakim.dillner@ki.se

Received May 20, 2013; **Accepted** July 22, 2013; **Published** July 29, 2013

Citation: Bzhalava D, Dillner J (2013) Bioinformatics for Viral Metagenomics. J Data Mining Genomics Proteomics 4: 134. doi:10.4172/2153-0602.1000134

Next generation sequencing technologies might produce exact and/or nearly duplicated reads due to PCR amplification, PCR errors and/or sequencing errors [19,20]. Identifying and removing these reads, a process also called de-duplication, can significantly reduce computational resources for downstream analysis and improve assembly. Presence of duplicated reads might also introduce overestimation of species abundance. On the other hand, duplicated reads might also include natural duplicates that by chance originate from the same start from the same genomic position [19,20]. Highly abundant species have a higher chance to have natural duplicates [20] and their removal might introduce bias towards underestimation of abundances [19]. The Cdhit-454 tool identifies and distinguishes artificial and natural duplicates in 454 pyrosequencing datasets [19]. CD-HIT-DUP tool identifies duplicates from single or paired Illumina reads [21].

To obtain the dataset that contains reads of interest, e.g. the virus-related reads for viral metagenomics, sequences that are not a target of the investigation need to be filtered out. This decreases the risk of misassemblieS [18], and also speeds up downstream analysis. NGS projects directed towards detecting of viral communities, generated from human samples subjected to whole genome amplification (WGA) may contain more than 70% of human-related reads unless there has been prior separation of viral capsids or shorter DNAs from long chromosomal DNA [10] (Table 1) and viral reads typically constitute less than 1% of reads (Table 1). With prior selection for viral nucleic acids, the human and bacterial related reads will still be the most commonly obtained reads, followed by sequences classified as "other" and "unknown" [10,11] (Table 1). Enrichment for viral particles by ultracentrifugation is helpful in the analysis of serum samples (Table 1), but has not been useful in the analysis of biopsies or skin swabs (Table 1). Bacterial sequences and sequences classified as "other" and "unknown" may also be present in negative control samples (water) after NGS sequencing [11] (Table 1), and it is therefore imperative that all metagenomic sequencing projects also include sequencing of negative control samples [11]. The background sequences found in water samples might be present due to the background reactivity of Phi29 polymerase reaction [22] or represent environmental contamination. However, water controls have so far been found to be uniformly negative for viral sequences [11] (Table 1).

To identify possible contaminant sequences as well as sequences that are not of interest, the NGS sequences need to be aligned against reference sequences. Different alignment software's are available for different sequencing platforms. There are hash table based softwares such as SSAHA2 [23] MAQ [24] and BFAST [25] as well as suffix/prefix tries based such as BWA-SW [26], SOAP2 [27] and Bowtie2 [27]. Hash table based algorithms require a large amount of operating memory, whereas suffix/prefix tries requires less computational resources.

Assembly

NGS technologies produce billions of short reads from random locations in the genome by oversampling it. Assembly algorithms, in the process called *de novo* assembly; reconstruct original genomes present in the sample by merging short genomic fragments into longer contiguous sequences ("contigs"). There are two main types of *de novo* assembly programs: Overlap/Layout/Consensus (OLC) assemblers, most widely applied to the longer reads such as MIRA and Celera Assembler's CABOG pipeline and *de Bruijn Graph* Assemblers, most widely applied to the shorter reads such as Euler [28], Velvet (www.ebi.ac.uk), ABySS [29], All Paths [30] and SOAP *de novo* (http://soap.genomics.org.cn/). The different assembly algorithms have been reviewed elsewhere [31-33].

The possibility always exists that assembly algorithms may construct erroneous "chimeric" sequences by the assembly of 2 different sequences from different organisms or species, a problem that may be particularly relevant for viral metagenomics where the bio-specimens may contain a multitude of related viral sequences. To validate assembly results, we suggest to use several assembly algorithms, as well as to perform a re-mapping of all singletons reads to assembled contigs [3,10].

Taxonomic Classification and Bining

Similarity based methods

Taxonomic classification or bining of metagenomic reads can be divided into similarity and non-similarity based methods. One of the most famous similarity-based taxonomic classifications is performed by NCBI BLAST searches where sequences are compared to known genomes. However, a large part of the sequencing reads from *de novo* sequencing projects are classified as unknown [10,11]. This can result from incompleteness of public sequence databases or drawbacks of NGS technologies such as short read lengths and sequencing errors. Because metagenomes might contain a large amount of sequences that have very distant homologs or even no homologs at all in public databases, we suggest that the use of BLASTn [34] nucleotide searches is suboptimal and that more sensitive algorithms, prone to identify more distant homologs may be preferable. One such possibility is to search against the protein database using BLASTx, or the tBLASTx algorithm, that translates query and reference nucleotide sequences in all six frames and then compares them to each other. Remote protein homologs can also be identified by exploring conserved protein domains using BLAST (such as deltablast [35]) or HMM-based (such

Sample type	FFPE[1] Biopsies		Fresh Frozen Biopsies		Skin Swabs					Serum		Negative water control
Pre-amplification treatment after WGA	E-Gel	-	E-gel	-	E-gel	UC[2]	-	-	-	UC[2]	UC[2]	-
Sequencing platform	GS FLX	GS FLX	GS FLX	GS FLX	GS FLX	GS FLX	GS FLX	Ion PGM 300bp kit	Ion PGM 400bp kit	GS Junior	GS Junior	GS FLX
Human	63.9	37.3	95.5	99.8	42.6	2.1	69.1	77.3	76.3	23.6	37.3	2.8
Bacteria	14.6	21.3	3.1	0.1	36.8	61.1	24.2	18.3	18.3	36.8	26.3	52.2
Virus	0.0	0.2	0.0	0.0	0.1	0.1	0.3	0.4	0.3	0.7	0.8	0.0
Other	11.7	10.2	0.5	0.0	17.1	31.5	2.2	1.3	1.0	33.1	20.5	15.5
Unknown	9.8	30.9	0.9	0.0	3.5	6.1	4.2	2.7	4.1	5.9	15.1	29.5

[1]Formalin Fixed Paraffin Embedded. [2]Ultracentrifugation.

Table 1: Typical taxonomic assignmentof NGS reads (percent). Summary of results in previous studies using different types of biospecimens, pretreatments and NGS platforms [10,11].

as HMM-FRAME [36]) alignment against Pfam [37], CDD [38] and TIGRFAM [39] databases.

Conventional BLAST-based search tools are extremely time consuming and may take days or even weeks to complete when large metagenomic datasets need to be compared against nucleotide or protein databases. Paracel Blast (Striking Development) is software that helps to save time by executing searches on multiple non-shared-memory processors simultaneously.

To classify sequences from alignment results several methods have been developed. One of the first and most frequently used is MEGAN [40]. In BLAST searches, sequences might have multiple matches and MEGAN finds the 'Lowest Common Ancestor' node of all matching sequences in the phylogenetic tree, which reduces the risk of false positive matches. However, MEGAN might produce false negative results by discarding sequences if they do not satisfy user-defined cutoffs. Because the size of genome is related to the number of reads in metagenomic samples, MEGAN is suboptimal for quantitative metagenomic analyses. This problem has been addressed by the development of the GAAS (Genome relative Abundance and Average Size) tool [41] that iteratively weights each reference genome for all matching reads and the number of reads is then normalized to the length of their genomes. GRAMMy (Genome Relative Abundance estimates based on Mixture Model theory) [42] is another useful tool that, compared to GAAS models, reads assignment ambiguities, genome size biases and read distributions along the genomes on a unified probabilistic framework [42]. However, both GAAS and GRAMMy estimate similarities from the alignment qualities of the reads to the reference genomes and not from the reference genomes directly. Thus, they are suboptimal in case there are highly similar genomes in the reference databases. The Genome Abundance Similarity Correction (GASiC) considers reference genome similarities to correct the observed abundances estimated via read alignments [43].

Composition based methods

Taxonomic classification methods that explore composition of genome such as GC content, codon or short oligomer (k-mers) usage are called composition-based methods. Their advantage is that they can be used for taxonomic classification of sequences that do not have any homologs or are highly divergent from sequences in public databases. Composition-based methods are computationally faster compared to similarity-based methods. However, they have lower accuracy and are very dependent on sequence length.

Composition-based methods can be divided into (1) assignment dependent: PhyloPythia [44] and Phymm [45], (2) hybryd: SPHINX [46] and PhymmBL [47] that combine similarity-based and composition-basedapproaches and (3) assignment independent: Metacluster3 [48] and Metacluster4 [49], TETRA [50], variants of SOMs [51], CompostBin (http://arxiv.org/abs/0708.3098) and AbundanceBin [52]. However, originally these methods were not designed for analyzing viral metagenomic datasets. Existing taxonomic profiling tools have problems to realistically profile and estimate abundances of viral sequences [53]. MGTAXA (http://mgtaxa.jcvi.org) is a composition-based tool that uses approach of the Phymm bacterial classifier [45] but is designed to predict the taxonomic placement of viral metagenomic sequences. Taxy-Pro tool [53] performs mixture model based analysis of protein signatures for taxonomic profiling and has good performance for estimating virus abundances in metagenomic datasets [53].

Genotype abundances, community diversity and structure

To estimate the number of different genotypes (richness) and their relative abundances (evenness) in a metagenomic sample, simple read counts may introduce biases, because longer genomes have a higher chance to be sequenced [41]. Another problem is that large parts of metagenomic sequences are classified as unknown, most probably because of shortcomings in the similarity-based taxonomic classification methods, which might result in biased diversity estimates.

Microbial community structure and their differences between different metagenomic samples can pinpoint the influences of patterns of microbial communities and among them presence of yet unknown microbes. Viral metagenome diversity and Community structure estimation pipelines mainly consist of generating contig spectra by tools like Circonspect (http://biome.sdsu.edu/circonspect), calculating average genome size by tools like GAAS [54,55], and using these two parameters to estimate biodiversity by PHACCS [56]. Genotype abundances and community diversity is estimated by the number of different genotypes in the sample, defined as richness (alpha diversity) and their relative abundances and distribution, defined as evenness (gamma diversity) among the metagenomic samples [57]. The analysis is based on the assumption that more abundant organisms will have longer and higher coverage contigs whereas less abundant organisms will have many small and low coverage contigs in the sample (alpha diversity) [57,58]. The gamma diversity uses the same assumption but estimates diversities among different metagenomic samples [57,58]. It assembles mixed sequences from metagenomic samples to be compared and the amount of similarity is measured by the degree of overlap (i.e., if fragments from one sample can be assembled with fragments from another sample) between the sequences from different samples. Monte Carlo analyses are then performed to estimate the degree of morphing [58].

Sequential blast analysis

Sequential blast analysis is another technique used to find shared and non-shared sequences between metagenomic samples [59]. If there are more than two metagenomic samples one is chosen randomly and is compared to a second randomly selected metagenome, which is used as a BLASTn database [59]. Applying user-defined cutoffs, the common sequences are identified and used as BLASTn database to be compared with a third randomly chosen database. The procedure continues until all metagenomic samples are compared. The entire pipeline may be repeated several times for different random orderings [59].

Discussion

As NGS technologies continue to develop rapidly, the metagenomics and viral metagenomics fields are expanding rapidly. NGS instruments generate large amounts of data that increase the demand on bioinformatics tools and algorithms. We have reviewed some of the most commonly used bioinformatics tools used to construct bioinformatics pipelines for viral metagenomic analysis.

One of the biggest challenges for bioinformatics analysis is taxonomic classification of NGS data as many of the sequences have no homologs in the public databases or are highly divergent, which is especially true for viral sequences [60]. Taxonomic classification by composition-based methods is in its infancy and very few methods have been developed for viral sequence classification and abundance estimations in metagenomic datasets [60]. MGTAXA (http://mgtaxa.jcvi.org) and Taxy-Pro [53] are particularly useful in this regard. As viruses are underrepresented in current genomic reference databases,

accurate and realistic estimation of the proportion of viral DNA in metagenomics is a great challenge [53]. Thus, further development of viral sequence classification and abundance estimations methods is essential.

Sequence quality checking, identification and removal of sequences of no interest as well as artificial duplicates are necessary steps to obtain as realistic datasets as possible that represent the sequences of interest (e.g. virus-related reads for viral metagenomics). This will decrease the risk form is assembly [18] and reduce the computational resources for downstream analysis. Different downstream analyses require different quality filtering methods [18].

Because the field of viral metagenomics is rapidly developing, both regarding the NGS technologies used and the bioinformatics tools applied, comparison of results from different studies is difficult and establishment of open access databases with metagenomics data also faces challenges in international comparability. We think that regular reviews of the best practices in the bioinformatics used in viral metagenomics, their advantages and shortcomings, are essential for the development of this important field.

References

1. Liu L, Li Y, Li S, Hu N, He Y, et al. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.
2. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434-439.
3. Meiring TL, Salimo AT, Coetzee B, Maree HJ, Moodley J, et al. (2012) Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits. Virol J 9: 164.
4. Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2: 3.
5. Foulongne V, Sauvage V, Hebert C, Dereure O, Cheval J, et al. (2012) Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One 7: e38499.
6. Feng H, Shuda M, Chang Y, Moore PS (2008) Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science 319: 1096-1100.
7. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, et al. (2008) Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog 4: e1000212.
8. Willner D, Haynes MR, Furlan M, Hanson N, Kirby B, et al. (2012) Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. Am J Respir Cell Mol Biol 46: 127-131.
9. Johansson H, Bzhalava D, Ekström J, Hultin E, Dillner J, et al. (2013) Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. Virology 440: 1-7.
10. Bzhalava D, Johansson H, Ekström J, Faust H, Möller B, et al. (2013) Unbiased approach for virus detection in skin lesions. PLoS One 8: e65953.
11. Bzhalava D, Ekström J, Lysholm F, Hultin E, Faust H, et al. (2012) Phylogenetically diverse TT virus viremia among pregnant women. Virology 432: 427-434.
12. Johansson H, Bzhalava D, Ekström J, Hultin E, Dillner J, et al. (2013) Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. Virology 440: 1-7.
13. Ekström J, Bzhalava D, Svenback D, Forslund O, Dillner J (2011) High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. Int J Cancer 129: 2643-2650.
14. Hemminki K, Dillner J (2009) Editorial. Int J Cancer 125: vii.
15. Moore PS, Chang Y (2010) Why do viruses cause cancer? Highlights of the first century of human tumour virology. Nat Rev Cancer 10: 878-889.
16. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214.
17. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194.
18. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, et al. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10: 57-59.
19. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11: 187.
20. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME J 3: 1314-1317.
21. Li W, Fu L, Niu B, Wu S, Wooley J (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. Brief Bioinform 13: 656-668.
22. Hutchison CA 3rd, Smith HO, Pfannkoch C, Venter JC (2005) Cell-free cloning using phi29 DNA polymerase. Proc Natl Acad Sci U S A 102: 17332-17336.
23. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11: 1725-1729.
24. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858.
25. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4: e7767.
26. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
27. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.
28. Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 98: 9748-9753.
29. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117-1123.
30. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, et al. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18: 810-820.
31. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327.
32. Narzisi G, Mishra B (2011) Comparing de novo genome assembly: the long and short of it. PLoS One 6: e19175.
33. Finotello F, Lavezzo E, Fontana P, Peruzzo D, Albiero A, et al. (2012) Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. Brief Bioinform 13: 269-280.
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
35. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, et al. (2012) Domain enhanced lookup time accelerated BLAST. Biol Direct 7: 12.
36. Zhang Y, Sun Y (2011) HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. BMC Bioinformatics 12: 198.
37. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211-222.
38. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37: D205-210.
39. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371-373.
40. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377-386.
41. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, et al. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5: e1000593.
42. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. PLoS One 6: e27992.

43. Lindner MS, Renard BY (2013) Metagenomic abundance estimation and diagnostic testing on species level. Nucleic Acids Res 41: e10.

44. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4: 63-72.

45. Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods 6: 673-676.

46. Mohammed MH, Ghosh TS, Singh NK, Mande SS (2011) SPHINX--an algorithm for taxonomic binning of metagenomic sequences. Bioinformatics 27: 22-30.

47. Brady A, Salzberg S (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods 8: 367.

48. Leung HC, Yiu SM, Yang B, Peng Y, Wang Y, et al. (2011) A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics 27: 1489-1495.

49. Wang Y, Leung HC, Yiu SM, Chin FY (2012) MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. J Comput Biol 19: 241-249.

50. Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5: 163.

51. Chan CK, Hsu AL, Halgamuge SK, Tang SL (2008) Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9: 215.

52. Wu YW, Ye Y (2011) A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. J Comput Biol 18: 523-534.

53. Klingenberg H, Aßhauer KP, Lingner T, Meinicke P (2013) Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. Bioinformatics 29: 973-980.

54. Angly F, Willner D, Prieto-Davó A, Edwards RA, Schmieder R (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5: e1000593.

55. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41.

56. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41.

57. Sun S, Chen J, Li W, Altintas I, Lin A, et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Res 39: D546-551.

58. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368.

59. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4: e7370.

60. Fancello L, Raoult D, Desnues C (2012) Computational tools for viral metagenomics and their application in clinical research. Virology 434: 162-174.

This article was originally published in a special issue, **Bioinformatics for Highthroughput Sequencing** handled by Editor: Dr. Heinz Ulli Weier, Lawrence Berkeley National Laboratory, USA