Computational Methods to Explore Viral Species from Wastewater Metagenomics Data.

Natàlia Timoneda[123], Joel Sabaniego[12], Xavier Fernandez-Cassi[3], Rosina Girones[3], Josep F Abril[12]

1 Departament de genètica, Universitat de Barcelona, Av Diagonal 643, 08028 Barcelona, Catalunya, Spain.
2 Institut de Biomedicina (IBUB), Universitat de Barcelona, Av Diagonal 643, 08028 Barcelona, Catalunya, Spain.
3 Departament de microbiologia, Universitat de Barcelona, Av Diagonal 643, 08028 Barcelona, Catalunya, Spain.

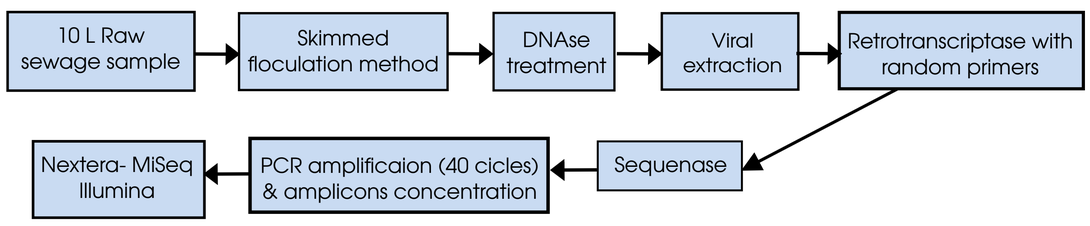
Summary

Treated wastewater is increasingly recognized as a resource of water nutrients and is reused in industry, for landscape irrigation, aquifer recharge and, especially in Mediterranean Europe, also for irrigation of fresh produce for human consumption. Wastewater contains many potential well known pathogenic bacteria and viruses; but also other potential emerging bacterial and viral pathogens largely unknown.

Although ICTV recognized 2243 viral species to date, the most recent estimates determine that we have identified and characterized less than 0.1% of the viruses present on this planet. However this number itself is likely a gross underestimate. For instance, approximately 40% of all diarrhea, being the third leading infectious cause of death worldwide, cases are unknown etiology.

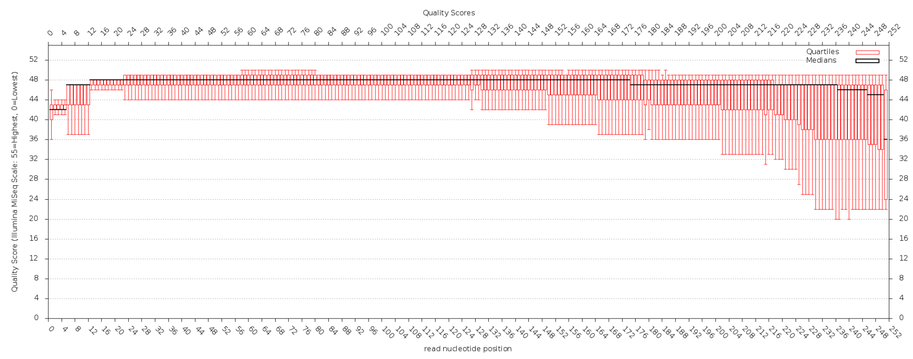
Our main objective is expand the list of known viruses present in urban sewage using metagenomics analyses. For that, we have generated a customized protocol to analyses viral wastewater data from NGS (MiSeq Illumina); and have had to adjust the parameters for the distinct bioinformatics tools used. Roughly; main steps of this protocol include standard cleaning of the raw metagenomics sequences, filtering most informative ones and performing search in order to detect known species. After that, unmapped sequences, at both levels raw reads and assembled contigs, may correspond to novel variants or species; so the sequences candidates are classified to facilitate the posterior experimental validation and markers selection.

1. Preparation sample for Illumina.



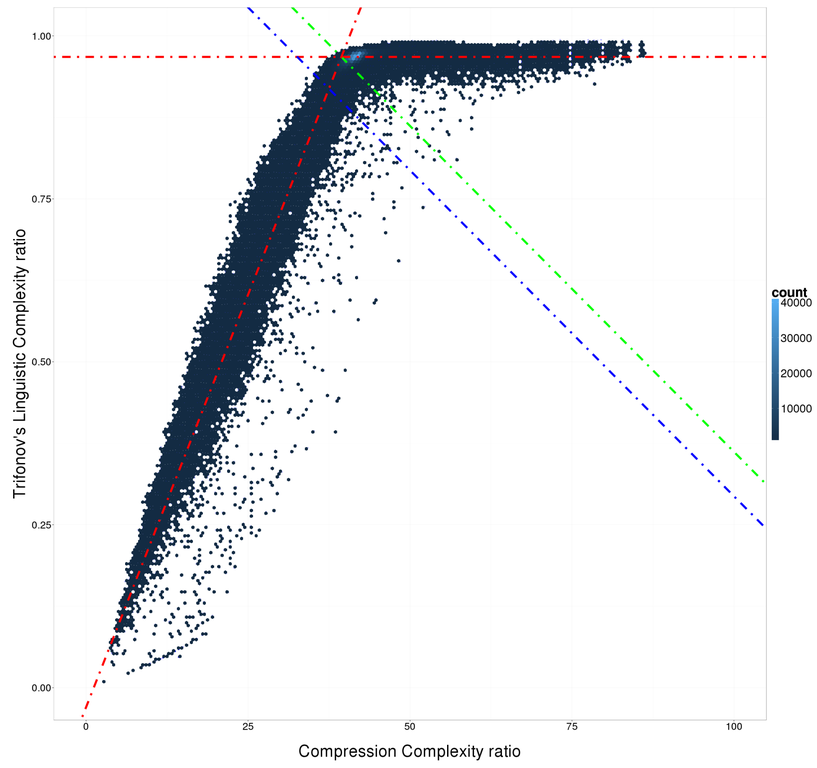
2. Pre-analysis of the samples, cleaning and trimming.

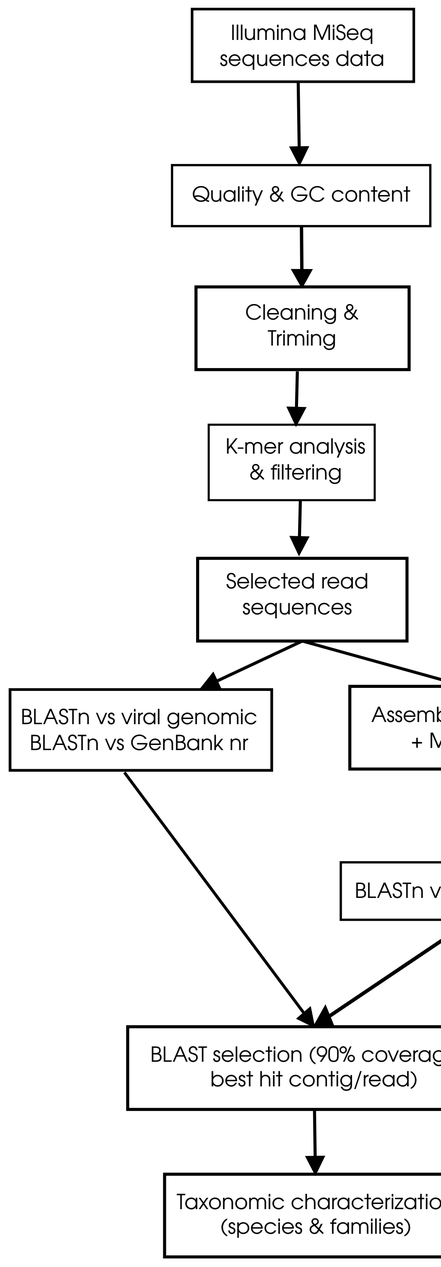
Raw sequences were processed with the FASTX software. To analyze the quality and nucleotide distribution statistics, to remove barcodes or noise, sequencing adapters/linkers,and finally sequences were trimmed based on



3. K-mer complexity.

To analyze the sequence complexity (mostly to remove repetitive sequences), Trifonov's linguistic complexity and compression ratio were calculated. Reads were filtered with a lineals model defined by a line 5% under the slope of 45º with respect the inflexion point of the parameters provided by the k-mer analysis.



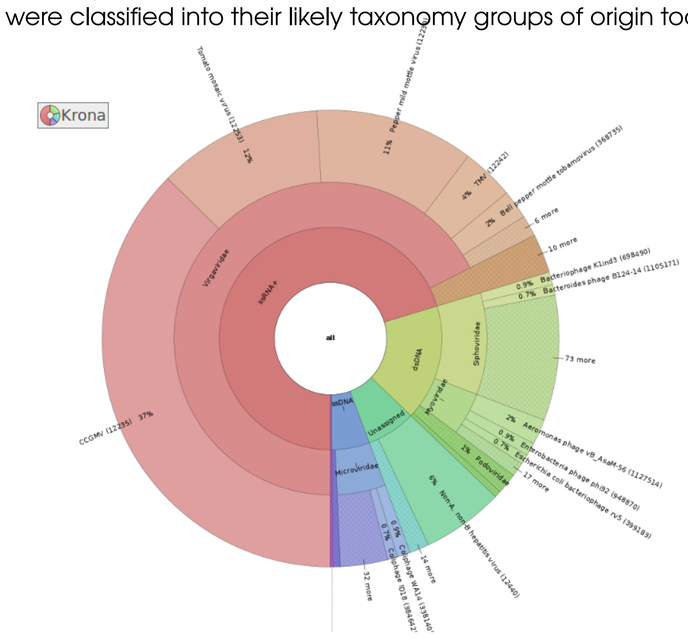


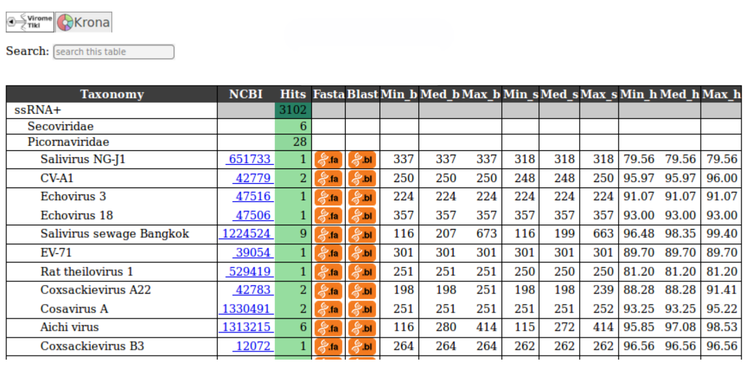
4.Sequence annotation.

After trimming, clipping and complexity filtering, sequences were sequentially compared against: (I) GenBank nr database using BLASTN; (II) viral and phage genome databases using BLASTN. Sequences with no significant hit were classified as "unassigned". On the basis of the best BLAST result and 90% coverage cutoff, sequences were classified into their likely taxonomic groups of origin.

5.Sequence assembly.

Reads that passed all the prepocessing steps were assembled into contigs using MetaVelvet (version 1.2.10). The singletons and contigs files were merged for each sample and a set of viral assembled sequences was created. Those sequences were annotated by BLASTN over viral and phage genomes database. Sequences without significant hits were classified as "unassigned".On the basis of the best BLAST result, sequences were classified into their likely taxonomy groups of origin too.



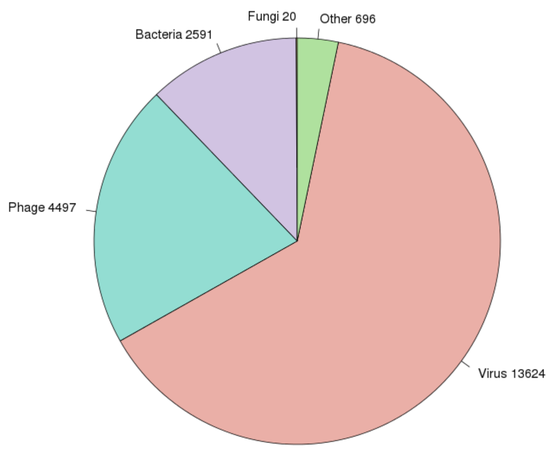


Search: search this table

Taxonomy	NCBI	Hits	Fasta	Blast	Min_b	Med_b	Max_b	Min_s	Med_s	Max_s	Min_h	Med_h	Max_h
ssRNA+		3102											
Secoviridae		6											
Picornaviridae		28											
Salivirus NG-J1	651733	1			337	337	337	318	318	318	79.56	79.56	79.56
CV-A1	42779	2			250	250	250	248	248	250	95.97	95.97	96.00
Echovirus 3	47516	1			224	224	224	224	224	224	91.07	91.07	91.07
Echovirus 18	47506	1			357	357	357	357	357	357	93.00	93.00	93.00
Salivirus sewage Bangkok	1224524	9			116	207	673	116	199	663	96.48	98.35	99.40
EV-71	39054	1			301	301	301	301	301	301	89.70	89.70	89.70
Rat theilovirus 1	529419	1			251	251	251	250	250	250	81.20	81.20	81.20
Coxsackievirus A22	42783	2			198	198	251	198	198	239	88.28	88.28	91.41
Cosavirus A	1330491	2			251	251	251	251	251	252	93.25	93.25	95.22
Aichi virus	1313215	6			116	280	414	115	272	414	95.85	97.08	98.53
Coxsackievirus B3	12072	1			264	264	264	262	262	262	96.56	96.56	96.56

6.Bowtie alignment and metagenome coverage analysis.

Filtered reads were aligned with Bowtie2 (version 2.2.2) against the viral genomes database and processed with samtools (version 0.1.18). Coverage ratio over genome was compared against local peaks to discard sequencing and assembly artifacts.



Acknowledgments: This work was supported by AGL2011-30461-C02-01/ALI (RG), BFU2009-09102 (JFA) (Ministerio de Ciencia e Innovación), 2009-SGR-001018 (JFA) (Generalitat de Catalunya) (JFA) and RecreCaixa 2012 "Improving deep sequencing techniques for fast identification of viral pathogens in water and Food". Spanish Ministry of Science and Innovation (grant number BES-2012-053084) (NT) and fellow of the Catalan Government "AGAUR" (FI-DGR) (grant number 2014FI_B1 00086) (XFC).