COMPARATIVE METAGENOMICS BY CROSS-ASSEMBLY

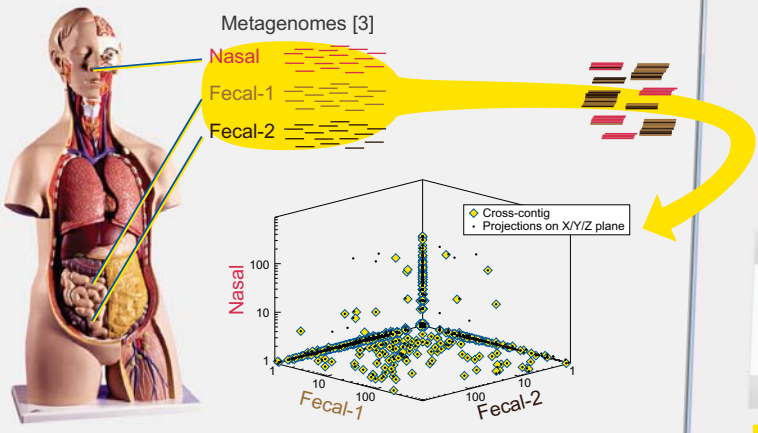
Bas E. Dutilh[1,2], Robert Schmieder[3], Jim Nulton[4], Ben Felts[4], Peter Salamon[4], Robert A. Edwards[3] and John L. Mokili[5]

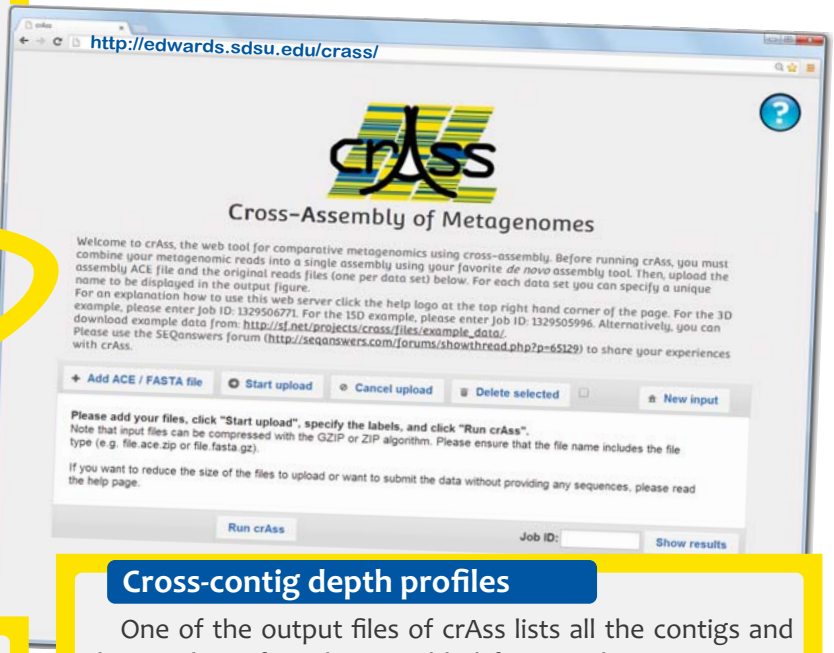
[1] Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud university medical centre, Nijmegen, The Netherlands. [2] Department of Marine Biology, Institute of Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. [3] Computational Science Research Center, [4] Department of Mathematics, and [5] Department of Biology, San Diego State University, San Diego, CA, USA.

Cross-assembly

Metagenomes are often interpreteded by mapping reads to an annotated reference database. Necessarily, search algorithms sacrifice either sensitivity or speed. Moreover, many microbes lack sequenced relatives in the database, resulting in large numbers of unknowns that are often ignored in further analyses [1].

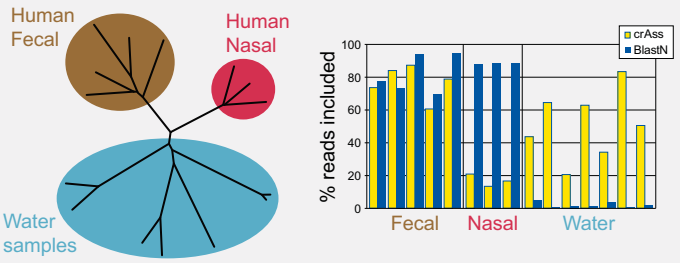
A promising alternative is reference-independent comparative metagenomics by cross-assembly [2].





Comparative metagenomics

The plot above shows many cross-contigs that exclusively contain reads from both fecal metagenomes (the yellow diamonds in the bottom plane), while few contigs combine reads from the fecal and nasal metagenomes. Thus, we can calculate similarities between metagenomes based on the cross-contigs and create a cladogram [2].



As cross-assembly is independent of a reference database, it allows relatively unexplored environments (e.g. water samples [4, 5]) to be compared comprehensively.

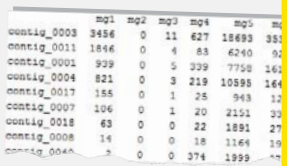
The impact of chimeras

Assembly of metagenomes has been challenged as the presence of sequencing reads from different strains and species may result in chimeras, i.e. contigs containing reads from different genomes [6]. Chimerization is more frequent for closely related species, and less severe for more distant organisms. Thus, we expect that chimeras are more likely to involve reads from similar biomes, rather than creating spurious links between unrelated metagenomes.

To address the impact of chimeras, we recommend to compare cross-assemblies with stringent/permissive assembly parameters, containing fewer/more chimeras, respectively. We find that these are often very similar.

Cross-contig depth profiles

One of the output files of crAss lists all the contigs and the number of reads assembled from each metagenome. Contigs with correlating occurrence profiles across metagenomes may be derived from the same genome [7] or otherwise related biologically (below).

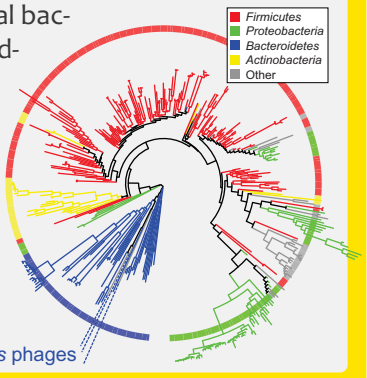


	mg1	mg2	mg3	mg4	mg5	mg
contig_0003	3456	0	11	627	18693	351
contig_0011	1846	0	4	83	6240	92
contig_0001	939	0	5	339	7750	162
contig_0004	821	0	3	219	10595	164
contig_0017	155	0	1	25	943	12
contig_0007	106	0	1	20	2151	33
contig_0018	63	0	0	22	1891	27
contig_0008	14	0	0	18	1164	19
contig_0040	3	0	0	374	1999	17

We identified cross-contigs with correlating depth profiles across twelve viral metagenomes derived from human feces. After careful reassembly, we recovered the genome of a novel bacteriophage that was not identified by mapping the metagenome to a database of known sequences.

Phage-host prediction

To predict the host of this novel bacteriophage, we screened 151 fecal metagenomes [8] for its presence, as well as the presence of 404 intestinal bacterial genomes that were considered as potential hosts. The presence profiles were correlated and clustered, and the phage clustered with its proposed *Bacteroidetes* host. Two previously known *Bacteroides* phages clustered similarly, providing a positive control.



References

[1] Mokili, Rohwer and Dutilh (2012) Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2**: 1–15.
[2] Dutilh *et al.* (2012) Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **28**: 3225–3231.
[3] Nakamura *et al.* (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One* **4**: e4219.
[4] Rosario *et al.* (2009) Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* **11**: 2806–2820.
[5] Lopez-Bueno *et al.* (2009) High diversity of the viral community from an Antarctic lake. *Science* **326**: 858–861.
[6] Pignatelli and Moya (2011) Evaluating the fidelity of *de novo* short read metagenomic assembly using simulated data. *PLoS ONE* **6**: e19984.
[7] Albertsen *et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* **31**: 533-538.
[8] Peterson *et al.* The NIH Human Microbiome Project. *Genome Res.* **19**: 2317-2323.