IMP: Imperial Metagenomics Pipeline for high-throughput sequence data

Lesley Hoyles[1,2] | James C. Abbott[1,3] | Elaine Holmes[1] | Jeremy K. Nicholson[1] | Marc-Emmanuel Dumas[1] | Sarah A. Butcher[1,3]

We have developed a modular pipeline (Figure 1) for analysis of metagenomic sequence data. The pipeline is undergoing validation using read data from a recent metagenomic study on the microbiome associated with liver cirrhosis [1].

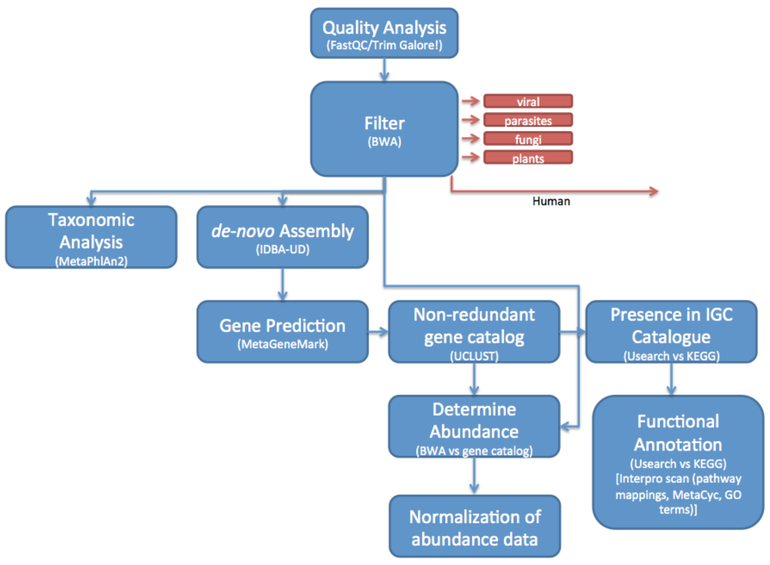
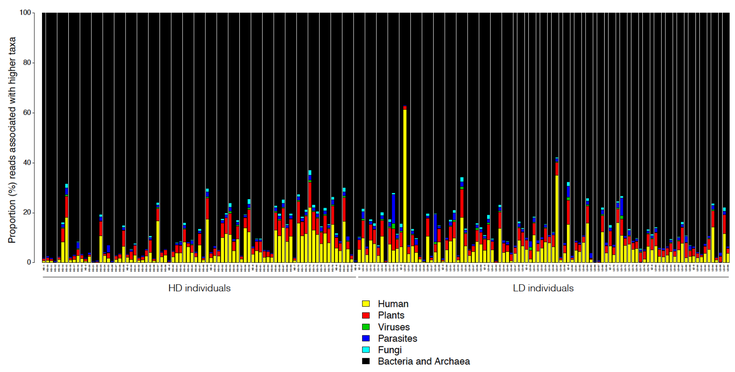


Figure 1. IMP: Imperial Metagenomics Pipeline. TrimGalore was used to quality trim data. Reads were filtered into human, viral, parasite, fungi, plant and bacterial/archaeal 'bins' using BWA [2] to map the reads against reference genomes, including the human genome, all available viral, protozoan and helminth, and plant genomes, and a selection of gut-associated fungi chosen based on the availability of whole genomes of species listed by [3]. Reads not assigned to any of the aforementioned groups were assumed to be of archaeal/bacterial origin. For the archaeal/bacterial data, taxonomic affiliations and abundance data were determined using MetaPhlAn 2.0 [4]. Bacterial/archaeal reads were assembled into contigs using IDBA-UD [5] in two phases – firstly on a per-sample basis, then using a pool of reads which remained unassembled following the first round of assembly. Putative genes were determined using MetaGeneMark [6,7], then clustered at the protein level using UCLUST 7.0.1090 [8] to create a non-redundant gene catalog. Functional classification of gene clusters was carried out by searching cluster centroids against a database of KEGG proteins (release 2015-05-1) with USEARCH [8], and functional domains/ pathway associations using InterProScan [9].

Sequence data for 83 healthy controls and 98 patients were processed (Figure 2). Human data were not analysed further because of ethical considerations. Plant, fungal and parasite data are currently presented on a presence/absence basis (Figure 5), but the modular nature of the pipeline means analyses of these datasets can be easily expanded in the future. Similarly, viral data are given on a presence/absence basis but are not presented graphically due to the complexity of the data (3092 virus groups represented across the patient samples).

Figure 2. Representation of higher taxa in sequence data.

	Raw read pairs	Trimmed read pairs
Range	12665510 – 109952378	10447268 – 101564583
Median	25158585	21806931
Average ± SD	29985569 ± 15092012	26139041 ± 13597198



MetaPhlAn 2.0 was used to generate abundance data (Figure 3).

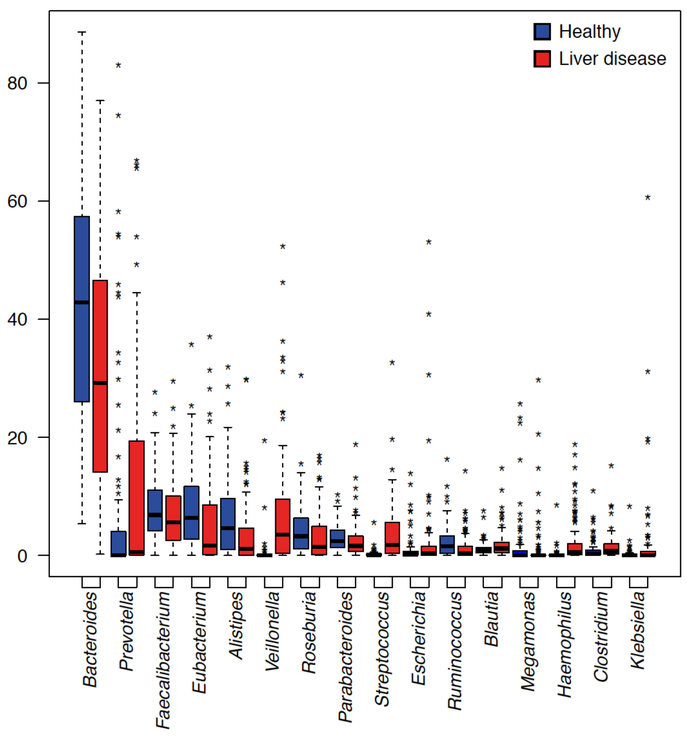


Figure 3. MetaPhlAn 2.0 analyses of metagenomic data from [1] using IMP. In agreement with [1], *Bacteroidetes* and *Firmicutes* represented the most abundant taxa in patient samples; *Veillonella*, *Streptococcus*, *Clostridium* and *Prevotella* were enriched in the liver-cirrhosis group; and *Eubacterium* and *Alistipes* were amongst the most dominant bacteria in the healthy controls.

Taxonomic data could be used to split patients into disease and healthy groups (Figure 4).

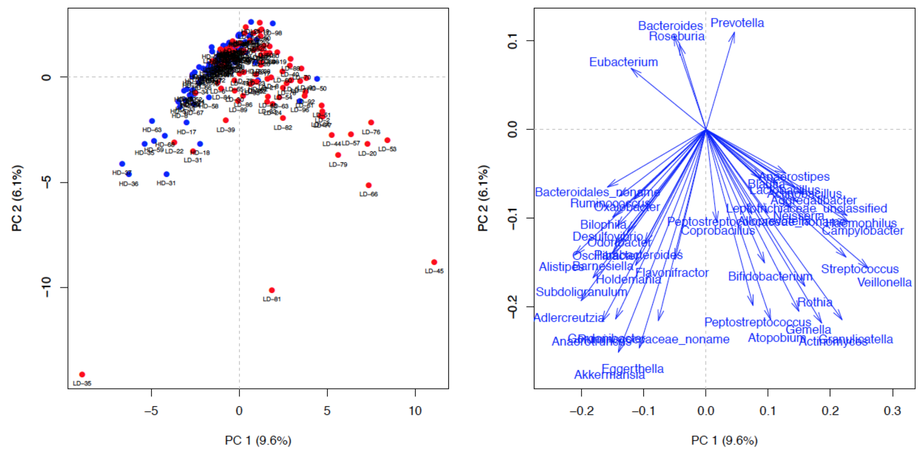


Figure 4. Principal component analysis of genus-level data generated from MetaPhlAn 2.0 outputs. After removal of three outliers from the liver-cirrhosis group, patients could be separated based on health status. (L) Scores plot; (R) loadings plot.

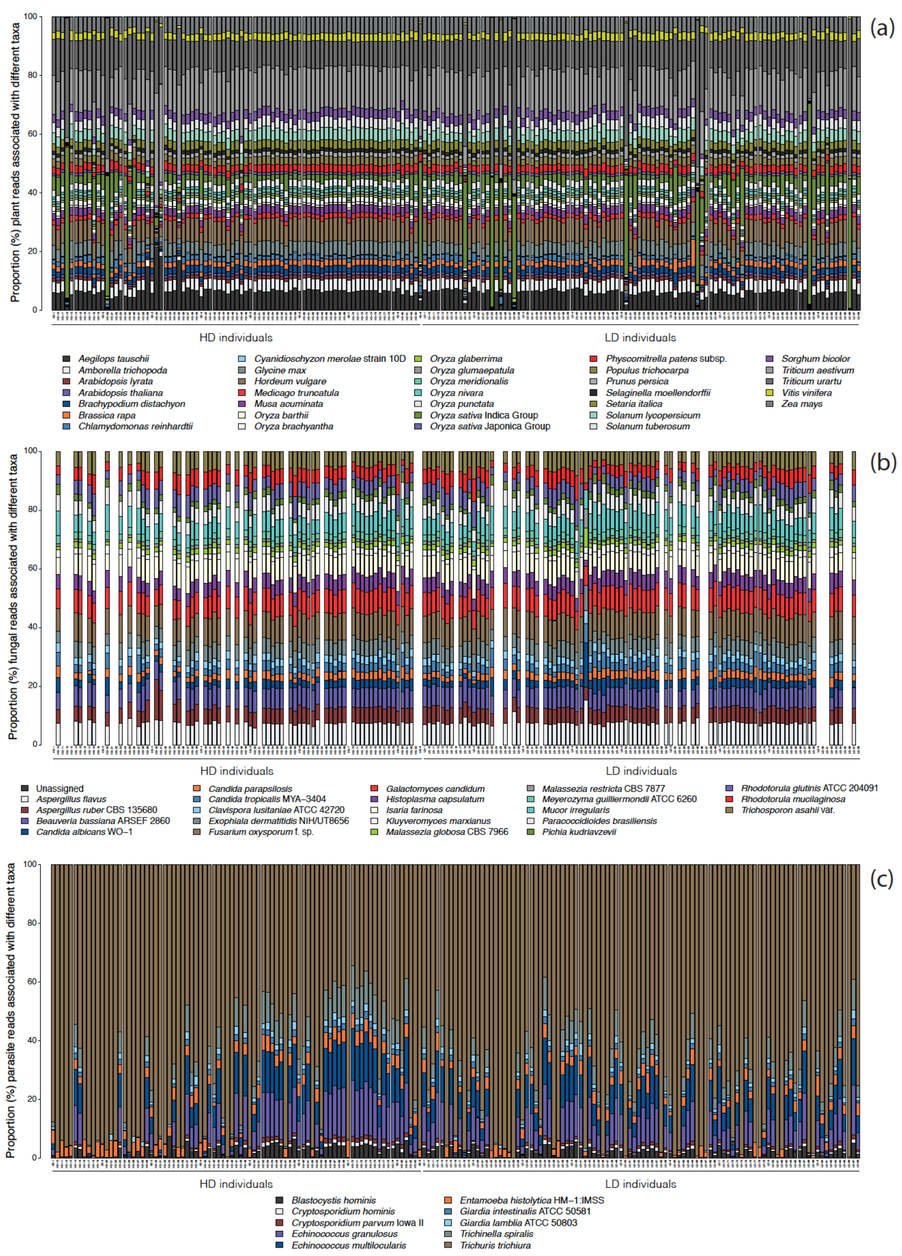


Figure 5. Assignment of reads from samples to (a) plant, (b) fungal or (c) parasite genomes. It is notable that several samples have no fungal DNA associated with them.

Selection of tools for use in the pipeline was made following assessment of numerous options; for example, a number of *de-novo* assemblers were assessed on both raw assembly statistics and a measure of potentially chimeric contigs produced, based upon the number of genera the reads associated with each contig originated from (Figure 6).

Assembler	Contigs	N50 (bp)	Max. length (bp)	Total length (bp)	MetaGeneMark predictions	MetaGeneMark N50
Velvet (k=61)	30086	1437	158607	38070000	55829	741
MetaVelvet (k=61,Training=Florinash)	30646	1405	158607	38220000	56446	732
MetaVelvet (k=61, Training=HumanGut)	30646	1405	158607	38220000	56446	732
IDBA-UD	57700	2859	250066	104300000	185774	774
Omega, l=60	10149	10022	122398	53680000	92389	684

	Genus contributing >5% of reads to contig					
Assembly	**1**	**2**	**3**	**4**	**5**	**6**
Velvet	95.79 %	2.77 %	0.89 %	0.44 %	0.11 %	0 %
MetaVelvet (Florinash)	95.76 %	2.93 %	0.76 %	0.38 %	0.11 %	0.05 %
MetaVelvet (Human Gut)	95.76 %	2.93 %	0.76 %	0.38 %	0.11 %	0.05 %
IDBA-UD	97.93 %	1.67 %	0.28 %	0.09 %	0.02 %	0 %
Omega	70.45 %	23.02 %	5.21 %	1.18 %	0.1 %	0.03 %

Figure 6. Results of comparison of *de-novo* metagenome assemblers

Although various tools have been selected for the different stages of the pipeline, the modular nature of the pipeline permits the ready replacement of these with alternatives should better or more appropriate methods become available in future.

Availability. Code is still being finalized, but will be available from www.ic.ac.uk/bioinfsupport/software when complete.

Author affiliations

[1]Division of Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London SW7 2AZ, United Kingdom

[2]Department of Biomedical Sciences, University of Westminster, 115 New Cavendish Street, London W1W 6UW, United Kingdom

[3]Centre for Integrative Systems Biology and Bioinformatics, Division of Molecular Biosciences, Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, United Kingdom

Contact Lesley Hoyles (lesley.hoyles11@imperial.ac.uk)

References

[1]Qin *et al.* (2014). *Nature* **513**, 59–64.

[2]Li & Durbin (2009). *Bioinforma Oxf Engl* **25**, 1754–1760.

[3]Gouba & Drancourt (2015). *Med Mal Infect* **45**, 9–16.

[4]Segata *et al.* (2012). *Nat Methods* **9**, 811–814 .

[5]Peng *et al.* (2012). *Bioinforma Oxf Engl* **28**, 1420–1428.

[6]Besemer & Borodovsky (1999). *Nucleic Acids Res* **27**, 3911–3920.

[7]Zhu *et al.* (2010). *Nucleic Acids Res* **38**, e132.

[8]Edgar (2010). *Bioinformatics* **26**, 2460–2461.

[9]Jones *et al.* (2014). *Bioinformatics* **30**, 1236–1240.

Imperial College London

FLORINASH contract N° Health-F2-2009-241913 is supported by funding under the Seventh Research Framework Programme of the European Commission