School of Biotechnology and Biomolecular Sciences

OTUreporter

amplicon to report in a single command

Xabier Vázquez-Campos[1], *David McFarlane*[2], *Angela M. Chilton*[3], *Jason Koval*[3], *Marc R.Wilkins*[1,3]

[1]NSW Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences,
[2]Research Technology Services, and
[3]Ramaciotti Centre for Genomics, UNSW, Sydney, NSW 2052, Australia

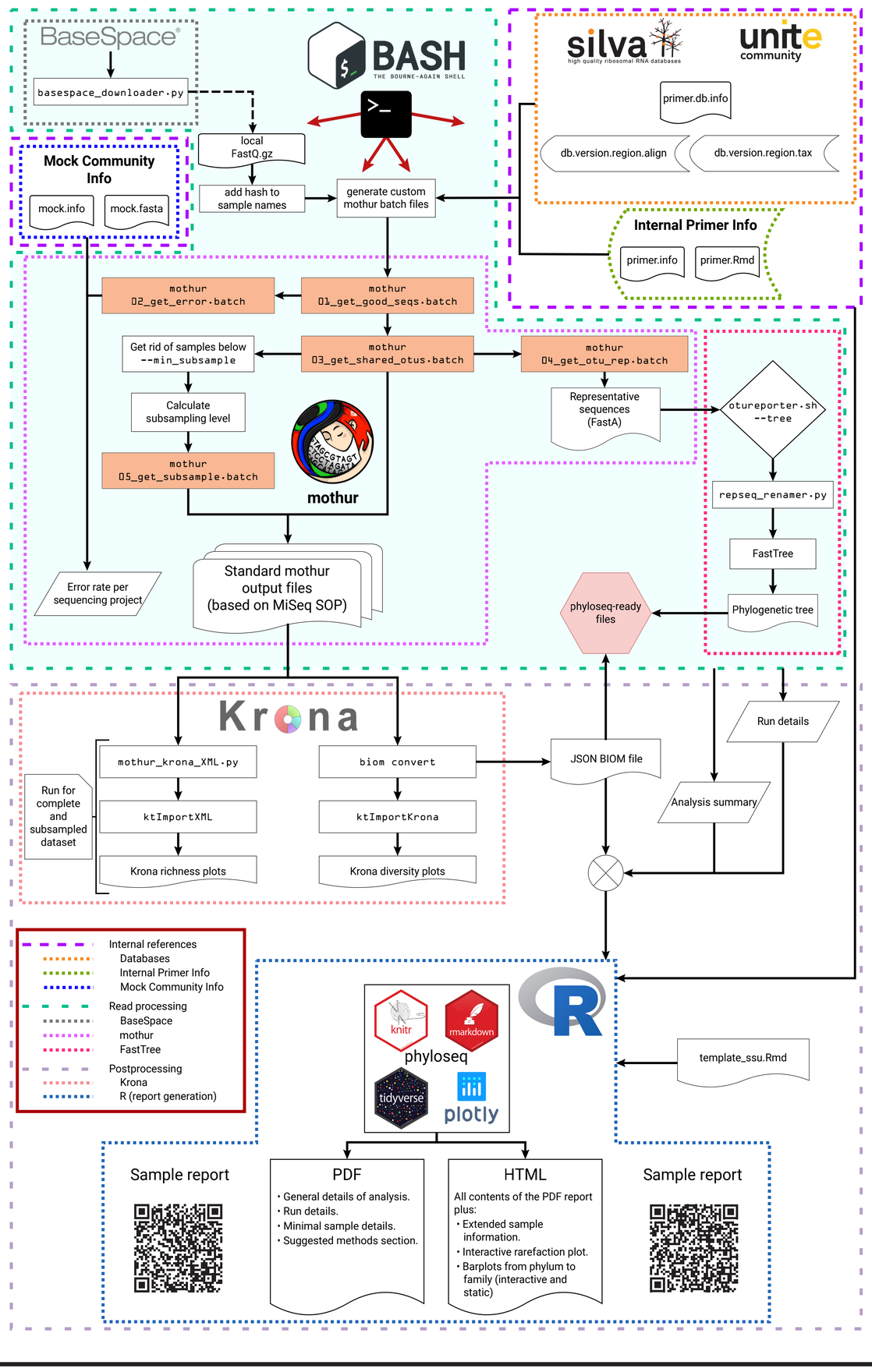
Introduction / Motivation

- Large increase in human and environmental microbiome amplicon samples over the last years (>10,000 in 2017) and still going up.
- Many new users: no idea how to process amplicon data and/ or they don't have the computational resources to analyse large datasets.
- Many comercial solutions are "black boxes": no idea what happens in the background. Not really used by scientists.

Objectives

- Generate a report for housekeeping and informative enough for customers and scientists with the *exact* parameters of the run.
- Provide a relatively easy to use pipeline that RAMAC personnel can use with minimal input/time spent to run it.
- Generate output files ready to use for downstream analysis
- Provide content that allows an initial exploration of the data.
- Able to deal with different amplicons, e.g. different genes or regions of the same gene.

Flow diagram

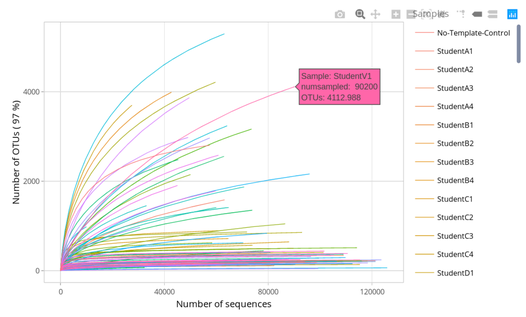


Sample output

General details of the analysis

Analysis parameter	Value
Run ID	full_dataset_out
Target lineage(s)	Universal
Excluded lineages	Chloroplast, Eukaryota, Mitochondria, unknown
Reference database	SILVA v132
Reference alignment	silva.v132.v4.align
Reference taxonomy	silva.v132.v4.tax
Mock community	Zymo
Mock reference file	zymo.ssu.fasta
OTU clustering cutoff	0.97
Minimum sequences	10000 sequences/sample
Subsampling level	10015 sequences/sample
Number of samples	94 (96 incl. controls)
Number of samples below threshold	5
'Bad' sample(s) ID(s)	No-Template-Control, StudentD4, StudentN1, StudentR3, StudentV2
Forward primer	515F-Y (5'-GTGYCAGCMGCCGCGGTAA-3')
Reverse primer	806RB (5'-GGACTACNVGGGTWTCTAAT-3')
Positive control(s)	Zymo-DNA-control
Negative control(s)	No-Template-Control
Tree generated	Yes

Rarefaction curve



Sample details

Sample[a]	RawReads[b]	NSeqFull[c]	SObsFull[d]	SObsSub[e]	CovFull (%)[f]	CovSub (%)[g]	InvSimpFull[h]	InvSimpSub[i]
StudentA1	82663	63203	1584	595.12±15.84	98.89	96.64±0.16	21.14	21.15±0.41
StudentA2	140057	121333	222	162.71±4.51	99.98	99.69±0.05	2.50	2.5±0.03
StudentA3	90277	82357	112	73.39±2.68	99.96	99.9±0.03	7.33	7.33±0.13
StudentA4	127041	109001	343	261.8±4.34	99.96	99.69±0.05	16.69	16.7±0.45
StudentB1	131785	108450	197	69.97±4.87	99.93	99.69±0.05	2.38	2.38±0.03
StudentB2	51605	42721	3989	2191.5±23.63	97.06	89.23±0.27	318.47	319.23±10.57
StudentB3	104037	89768	500	385.72±6.38	99.94	99.39±0.07	25.23	25.25±0.51
StudentB4	135742	118273	218	161.01±3.65	99.97	99.8±0.04	9.13	9.12±0.2
StudentC1	25684	22302	96	79.49±2.93	99.89	99.83±0.03	8.90	8.89±0.15
StudentC2	78588	64690	719	529.05±8.45	99.87	98.7±0.09	32.61	32.63±0.73
StudentC3	112745	88048	645	339.9±8.27	99.81	98.94±0.09	12.24	12.25±0.25
StudentC4	72398	61039	893	662.03±9.99	99.85	98.3±0.11	29.87	29.9±0.57
StudentD1	33782	27531	3700	2379.68±22.92	95.17	88.02±0.27	271.22	271.95±7.64

[a] Sample name derived from fastq.gz filename.
[b] Number of raw reads or read pairs.
[c] Number of sequences passing QC and chimera removal.
[d] Number of observed OTUs.
[e] Number of observed OTUs after subsampling.
[f] Good's coverage. Estimates what percentage of the species in a system (OTUs here) is represented in a sample.
[g] Good's coverage estimation after subsampling.
[h] Inverse Simpson Index. Diversity index less prone to bias due to unequal sampling efforts.
[i] Inverse Simpson Index after subsampling.

Discarded samples summary

Any sample that did not have a minimum of 10000 sequences after filtering, was not taken into account for calculating the subsampling level.

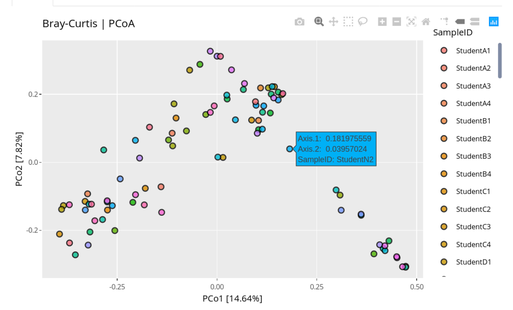
The following table contains details about the samples that did not achieve a minimum of 10000 sequences:

SampleName	RawReads	NSeqs	SObs	Coverage (%)	InvSimp
No-Template-Control	228	63	28	73.02	11.42
StudentD4	1088	862	380	76.68	167.84
StudentN1	150	57	43	38.60	69.39
StudentR3	87536	522	105	89.08	12.03
StudentV2	1273	1056	45	96.78	2.20

Bray-Curtis dissimilarity

The Bray-Curtis dissimilarity metric informs about how different samples are to each other. Unlike other measures, Bray-Curtis is better at handling sparse matrices (i.e. tables with a large proportion of zeroes) commonly found in ecological data sets (e.g. OTU tables).

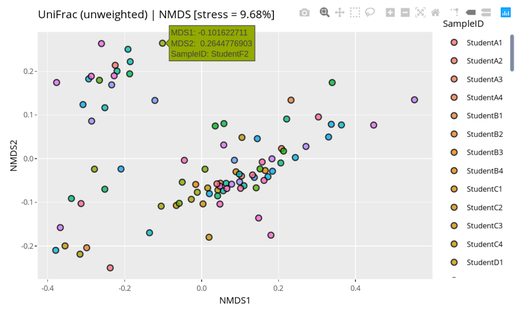
PCoA | NMDS



UniFrac (unweighted)

The UniFrac method measures the phylogenetic distance between samples based on the phylogeny of the constituent taxa. UniFrac uses the fraction of the branch length of the tree that leads to taxa that can be used to discriminate between samples. The original, ***unweighted UniFrac*** is a qualitative method (presence/absence of taxa) and might place too much weight on rare taxa.

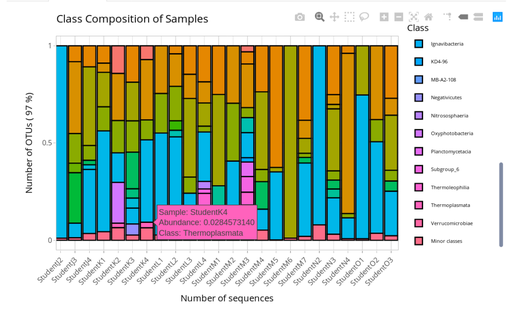
PCoA | NMDS



Class-level composition

Minor classes group all classes that did not reach a minimum of 2% of the whole community in any sample.

Static | Interactive



Contact

Xabier Vázquez-Campos
Email: xvazquezc@gmail.com
Twitter: @XabiVC
Bitbucket repository: xvazquezc/otureporter

Get a copy of the poster here:

Acknowledgements

Ignatius Pang for sharing his expertise with R, Åsa Pérez-Bercoff for providing the `repseq_renamer.py` script, and Gene Hart-Smith for the logo design.

Jai J. Tree and all staff from the MICR2011 Microbiology course at UNSW that provided feedback to fix several bugs and.