Metagenomics 101

Session 8: From genes and reads to profiles

Anna Heintz-Buschart May 2022



Metagenomics (+ other omics) pipeline



MP3



imp3.readthedocs.io

Metagenomics (+ other omics) pipeline



imp3.readthedocs.io





Repetition:

- ✤ mapping
- \clubsuit annotation

Gene abundance measures - functional profile calculation

- reads per gene / reads per function
- ✤ reads per kilobase
- copies per million
- ✤ average depth of coverage

Working with functional profiles

Mapping reads



Mapping reads





Sequence alignments (SAM format)

re	ad name	flag	referend	e ,	qua	lity p	artn	er	~2	alignment length read sequence
	SRR5947855.214477	99	test_contig_1	1	60	70M31S	=	1	69	AAGCAGTGACTATGTAGTCATCCAAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCTGTC
	:test contia 94,14,-	CC <mark>CACGAG</mark> -,38M63S,7,	(CCFFDDFHHHGHGIJ 0: RG:Z:test	: J@HI	GI IJJJ	IIJJJJGIII		JJGHIIJJ	ITOSDUT	JUARUNDENNUGENDUDUDUDUDUDUDUDU NNH.1.0 MU.2.70 A3.1.70 A3.1.0 3A.2
	SRR5947855.214477	147	test_contig_1	1	60	32S69M	=	1	-69	CGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGGACTCACTATCTAGTCATCCAAAGACAATGAAATAGCGA CIStance to
	GAAGGAATCCAGAAGATATT	TCCAGGTG			IGJIJIHE(GJJJJJJIJJJJ	IIGGEIJ	JIHGIIJJJ	IIIHIIJI	IIHGIIHBIJJJIGHEDCIDHGJJHHBHFFFFF@@@ NM:1:0 ME:Z:69 AS:1:69 XS:1:0 SA:Z

reterence SRR5947855.32<u>8367</u> 99 60 101M 136 TCCAAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGATTTGCGGAATGATGACCTTTC test contig 1 21 216 TAGTGCTTGGGCGGGCGTTGCATTTAA NM:i:0 MD:Z:101 AS:i:101 XS:i:0 RG:Z:test mismatching SRR5947855.326355 99 test_contig_1 24 60 101M 185 262 AAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGT0CGATTTGCGGA/TGATGACCTTTCTTAG = CCCFFFFFHGHGHJJJJJJJJEBEIJJJJJJJJEBEGHIIGIJGIGJIIIIJJIGFF>EHHFFFFEEEEDDCCCDDDDDDDDDDDDC9>BDDDEDEACDC AS:i:101 TGCTTGGGCGGGCGTTGCATTTAAATC NM:i:0 MD:Z:101

XS:i:0 RG:Z:test positions GACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGATTTGCGGAATGACCTTTCTTAGTGC SRR5947855.10987 99 test_contig_1 27 60 101M 125 199 = TTGGGCGGGCGTTGCATTTAAATCTAA CCCFFFFBFFHHGJJJIJIIJIIJJIFJJJJJGHIIJIIJJFJHHIIFHIJJJJJIIIACCE>AHEDFFFFFECEED>>@BABDBBDDB@@ACDADCCDDDC> NM:i:0 MD:Z:101 AS:i:101 XS:i:0 RG:Z:test SRR5947855.362726 163 test_contig_1 91 <u>TTCTTAGTGCTTGGGCGG</u>GCGTTGCATTTAAATCTAATGCAGCTTCATATAATTCTGGATT_ALLIGIIGIAA 60 101M 141 151 TTTGGCTTTACGTTCAAATAAACGAAA NM:i:0 MD:Z:101 AS:i:101 XS:i:0 RG:7:test 1.

55.10987 147	test_contig_1	125	60	101M	=	27	-199		
CAATGTCCTGCTAATAAAG	DDCDDCDCCDDDFEDE	EFFFEEHE	ннннлэг	IIIGEHIJ	JJIGHIH	GB?DIIIGJ	IGCHHFHEJ	JJJIHEJIJIJIJIHFHHIJJJJIHF>HHFFFDFCC@ NM:i:0 MD:Z:101 AS:i:101	
XS:i:0 RG:Z:test 55.328367 147	test contig 1	136	60	101M	=	21	-216		

AS:i:101 NM:i:0 MD:Z:101 TGTCCTGCTAATAAAGTCCAGTATGAG XS:i:0 RG:Z:test SRR5947855.362726 83 test_contig_1 141 60 101M 91 -151 AATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAGATAAAACACAATGTCC = TGCTAATAAAGTCCAGTATGAGAAGAG NM:i:0 MD:Z:101 AS:i:101 XS:i:0 RG:Z:test SRR5947855.326355 147 test_contig_1 185 60 101M 24 -262 AACGAAAGGCATAGATAAAACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAAATTTCTAAATATGA = TATTACCTGCAAAAACAAACATAACTG NM:i:0 MD:Z:101 AS:i:101 XS:i:0 RG:Z:test SRR5947855.343465 99 60 101M 257 154 ACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAATTTCTAAATATGATATTACCTGCAAAAAACAAA test_contig_1 204 =

CATAACTGTCAATGATAAAGCCATGAC AS:i:101 CCCFFFFFHHGHHJJJJJJJJJJJHHIJGEEGFGGGHIIJIIJJJFIJIIFIJJJIJIJJJJJJJJJJJGFHHGGEFFFFFFEEEE=CDDDDD NM:i:0 MD:Z:101 XS:i:0 RG:Z:test SRR5947855.285811 60 101M 134 GATATTACCTGCAAAAACAAACATAACTGTCAATGATAAAGCCATGACGACTTCAGCATTTACAACAGGAATTT 99 test_contig_1 257 = 290 AS:i:101 NM:i:0 MD:Z:101

CIGAR position of partner

GAGTCATTCCTTCGATAACAGTTTGAG XS:i:0 RG:Z:test SRR5947855.343465 147 GAGTCATTCCTTCGATAACAGTTTGAG DFDCCC?DDEDDDEEFFFFFHHHHHHHIJJJJJJJJJJJJJJJJJIIJII XS:i:0 RG:Z:test

SRR59478 ATAAAACA SRR59478

Jniversity of Amsterdam Life Sciences

AS:i:101

AACATAACTOTOAATCATAAAGCCATGACGACTTCAGCATTTACAACAGGAATTT

NM:i:0 MD:Z:101

read quality

Sequence alignments (SAM format)

St			
Bit		Bit	Description
	1	0x1	template having multiple segments in sequencing
	2	0x2	each segment properly aligned according to the aligner
	4	0x4	segment unmapped
	8	0x8	next segment in the template unmapped
	16	0x10	SEQ being reverse complemented
	32	0x20	SEQ of the next segment in the template being reverse complemented
	64	0x40	the first segment in the template
	128	0x80	the last segment in the template
	256	0x100	secondary alignment
	512	0x200	not passing filters, such as platform/vendor quality controls
	1024	0x400	PCR or optical duplicate
-	2048	0x800	supplementary alignment



Annotation of protein-coding genes

- positions on the contigs
- direction on the contigs
- translation
- information on completeness

.gff General feature format:

contig	source	type	start	enc	b	stra	nd	attributes
ontig_1001	Prodigal_v2.6	.3 CDS	3	479		+	0	ID=GGBJBNCP_01295;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNCP_01295;partial=11
ontig_1002	Prodigal_v2.6	.3 CDS	3	335		_	0	ID=GGBJBNCP_01296; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01296; partial=11
ontig_1003	Prodigal_v2.6	.3 CDS	1	387		+	0	ID=GGBJBNCP_01297; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01297; partial=11
ontig_1004	Prodigal_v2.6	.3 CDS	1	1053		-	0	ID=GGBJBNCP_01298; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01298; partial=11
ontig_1005	Prodigal_v2.6	.3 CDS	2	355		-	0	ID=GGBJBNCP_01299; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01299; partial=11
ontig_1006	Prodigal_v2.6	.3 CDS	3	473		+	0	ID=GGBJBNCP_01300; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01300; partial=11
ontig_1007	Prodigal_v2.6	.3 CDS	1	849		-	0	ID=GGBJBNCP_01301; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01301; partial=11
ontig_1008	Prodigal_v2.6	.3 CDS	67	303		+	0	ID=GGBJBNCP_01302; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01302; partial=01
ontig_1009	Prodigal_v2.6	.3 CDS	1	102		+	0	ID=GGBJBNCP_01303; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_01303; partial=10
ontig_100	Prodigal_v2.6	.3 CDS	2	628		-	0	ID=GGBJBNCP_00117; inference=ab initio prediction: Prodigal_v2.6.3; locus_tag=GGBJBNCP_00117; partial=10
					score		phase	



Functional databases

Pfam

Curated families/ontologies

- Pfam
- KEGG
- EggNOG



Large collections





Specialized databases

- antibiotics resistance: Resfams, CARD, ...
- specific metabolism: antiSMASH, CAZy, ...
- taxonomic/phylogenetic markers: BUSCO, CheckM, mOTUs, ...
- others: virulence, effectors, toxins, plasmids, phages, CRISPR...





Counting reads



Reads per gene

- e.g. featureCounts (subread) software
- count reads that map to positions overlapping with coding sequences
 - ➤ multi-overlap
 - Iength of overlap / overhang
 - > stranding (direction)
 - ➤ forward- and reverse read





12

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/05_counting_reads.htm



Reads per annotated functional class

- e.g. featureCounts (subread) software
- count reads that map to positions overlapping with annotated sequences
 - ≻ multi-overlap
 - Iength of overlap / overhang
 - ➤ stranding (direction)
 - ➤ forward- and reverse read
- or sum up reads from gene-level profile at class level



UNIVERSITY OF AMSTERDAM Life Sciences

Reads-level analysis

- profiles need to be compared across samples:
 - reads-per-gene analyses: for catalogue-based analyses (session 11)
 - reads-per-class analyses to compare samples
- > normalization to different sampling depths (mapping rates)
- count data for DESeq2 and related methods





Reads per gene/class per kb

- e.g. HUMAnN
- count reads that map to gene/class and divide by total length of gene/class

$$RPK_{gene} = \frac{reads \ mapped \ to \ gene \times 10^3}{total \ gene \ length}$$

different result for 1000 reads mapping to 50 genes than 1000 reads mapping to 5 genes

"Copies" per million

- e.g. HUMAnN
- normalize RPK to total number of reads to compare samples

$$RPK_{gene} = \frac{reads \ mapped \ to \ gene \times 10^3}{total \ gene \ length}$$
$$RPK_{gene} \times 10^6$$

$$CPM_{gene} = \frac{RPK_{gene} \times 10^6}{\sum RPK}$$

CPM vs RPKM





"Copies" per million

In my opinion, there is no good way to do a DE analysis of RNA-seq data starting from the TPM values. TPMs just throw away too much information about the original count sizes. Sorry, but I'm not willing to make any recommendations, except to dissuade people from thinking that TPMs are an adequate summary of an RNA-seq experiment.

Note that it is not possible to create a DGEList object or CPM values from TPMs, so trying to use code designed for these sort of objects will be counter-productive.

I see that some people in the literature have done limma analyses of the log(TPM+1) values and, horrible though that is, I can't actually think of anything better, given TPMs and existing software. One could make this a little better by using eBayes with trend=TRUE and by using arrayWeights() to try to partially recover the library sizes. Please do not take that as a recommendation though!

ADD COMMENT • link

➢ no good statistics yet



Average depth of coverage

- related to RPK
- useful, if length of reads is important
- use for within-sample analyses





Recap

- reads per function/class
 - ➤ straight-forward
 - counts of reads matching to function
 - need normalization
 - statistical methods are developed (e.g. DESeq2)

- "copies" per million
 - ➤ based on reads per function,
 - but normalized to gene length
 - normalized to sampling depth
 - statistical models?
- average depth of coverage
 based on numbers of reads covering each position
 - mostly useful for within-sample comparisons



Thanks for your attention!



a.u.s.heintzbuschart@uva.nl

SP C2.205



github.com/a-h-b



twitter.com/_a_h_b_

