

Metagenomics 101

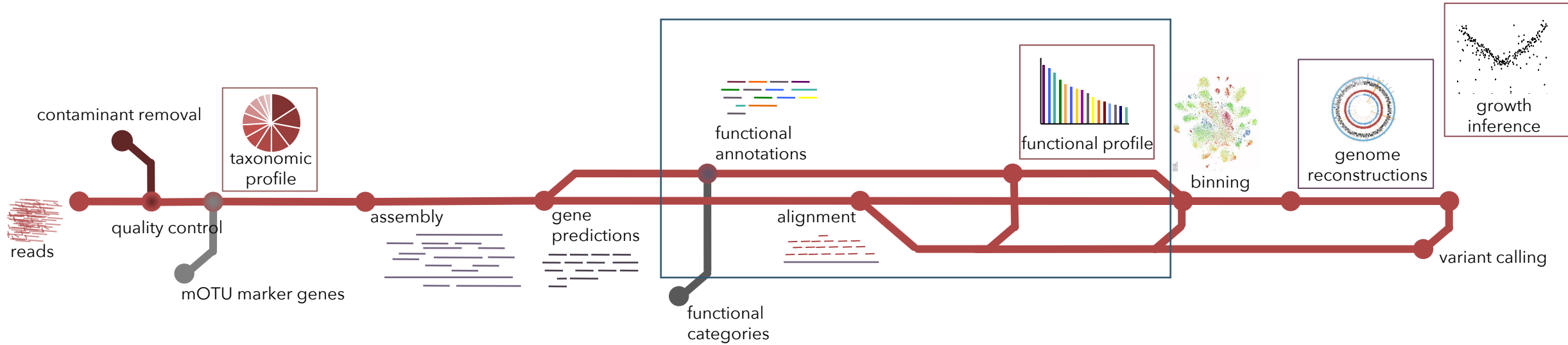
Session 8: From genes and reads to profiles

Anna Heintz-Buschart

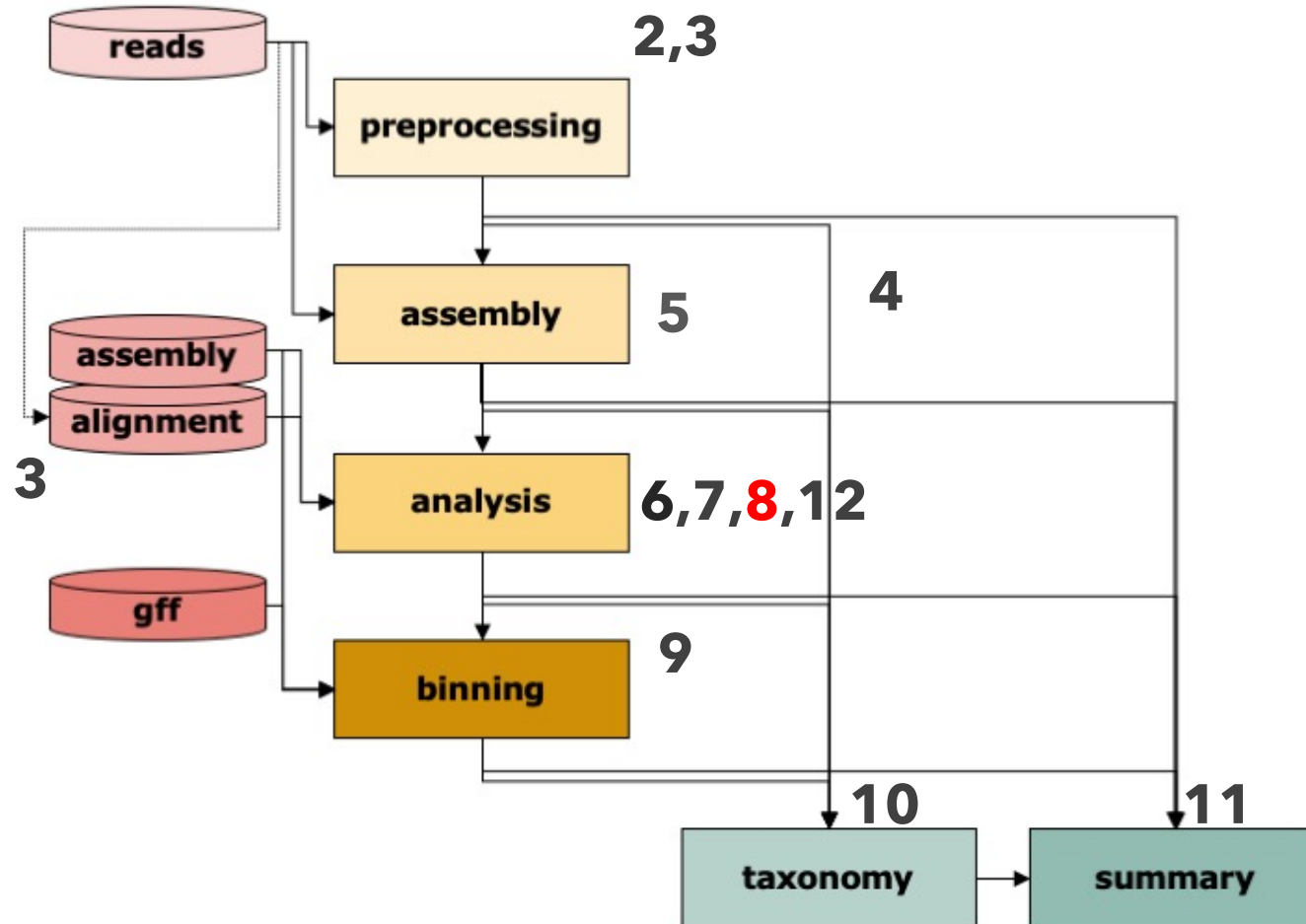
May 2022



Metagenomics (+ other omics) pipeline



Metagenomics (+ other omics) pipeline



Today

Repetition:

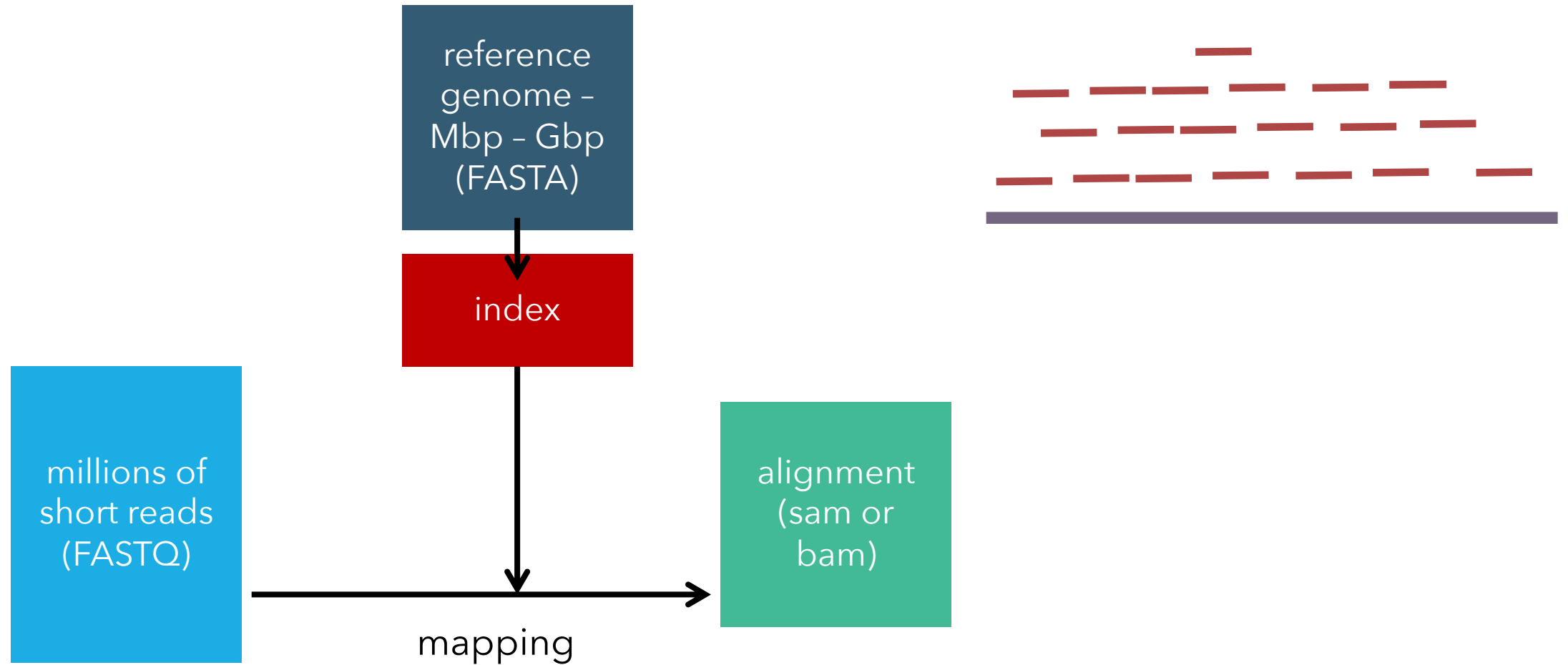
- ❖ mapping
- ❖ annotation

Gene abundance measures - functional profile calculation

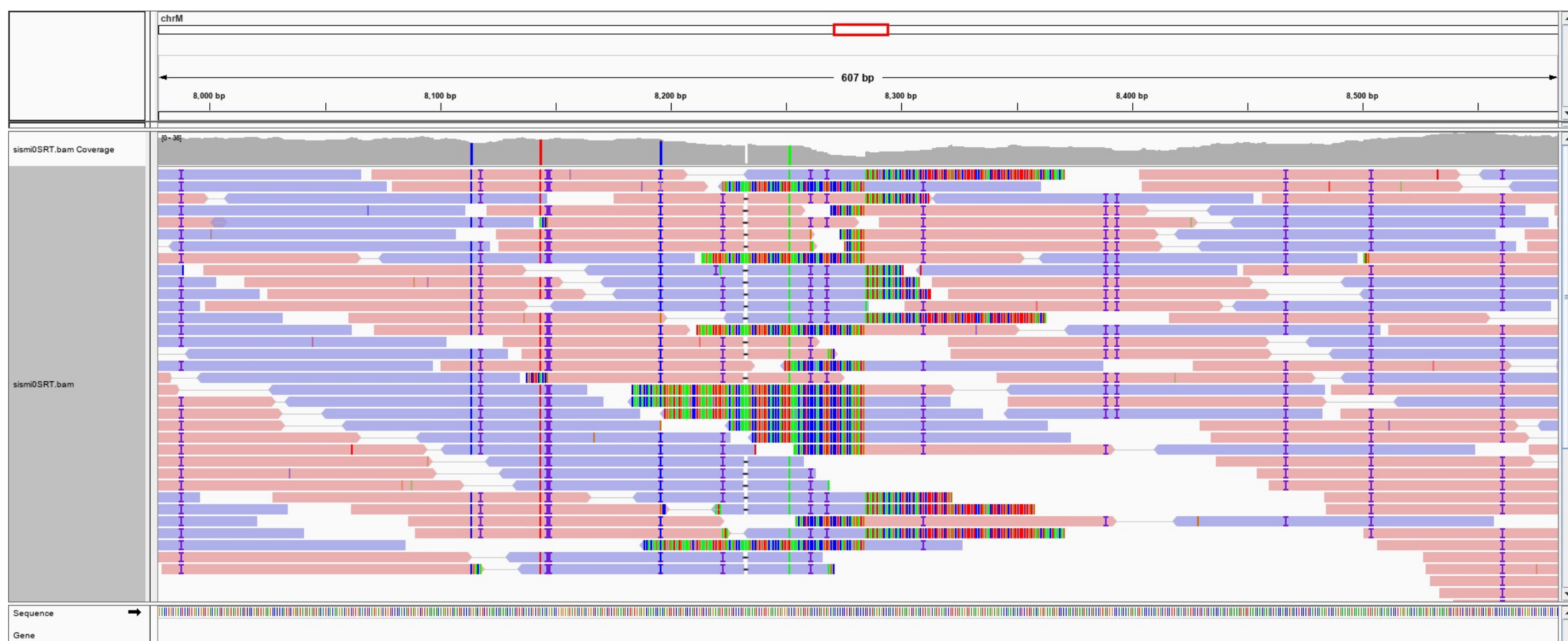
- ❖ reads per gene / reads per function
- ❖ reads per kilobase
- ❖ copies per million
- ❖ average depth of coverage

Working with functional profiles

Mapping reads



Mapping reads



Sequence alignments (SAM format)

read name	flag	reference	position	quality	name of read	partner	~alignment length	read sequence
-----------	------	-----------	----------	---------	--------------	---------	-------------------	---------------

[illegible]

distance to
reference

mismatching
positions

alignment
score

CIGAR position of partner

read quality

Sequence alignments (SAM format)

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	end	strand	attributes
contig_1001	Prodigal_v2.6.3	CDS	3	479	.	+ 0 ID=GGBJBNC_01295;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01295;partial=11
contig_1002	Prodigal_v2.6.3	CDS	3	335	.	- 0 ID=GGBJBNC_01296;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01296;partial=11
contig_1003	Prodigal_v2.6.3	CDS	1	387	.	+ 0 ID=GGBJBNC_01297;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01297;partial=11
contig_1004	Prodigal_v2.6.3	CDS	1	1053	.	- 0 ID=GGBJBNC_01298;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01298;partial=11
contig_1005	Prodigal_v2.6.3	CDS	2	355	.	- 0 ID=GGBJBNC_01299;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01299;partial=11
contig_1006	Prodigal_v2.6.3	CDS	3	473	.	+ 0 ID=GGBJBNC_01300;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01300;partial=11
contig_1007	Prodigal_v2.6.3	CDS	1	849	.	- 0 ID=GGBJBNC_01301;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01301;partial=11
contig_1008	Prodigal_v2.6.3	CDS	67	303	.	+ 0 ID=GGBJBNC_01302;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01302;partial=01
contig_1009	Prodigal_v2.6.3	CDS	1	102	.	+ 0 ID=GGBJBNC_01303;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_01303;partial=10
contig_100	Prodigal_v2.6.3	CDS	2	628	.	- 0 ID=GGBJBNC_00117;inference=ab initio prediction:Prodigal_v2.6.3;locus_tag=GGBJBNC_00117;partial=10

score phase

Functional databases

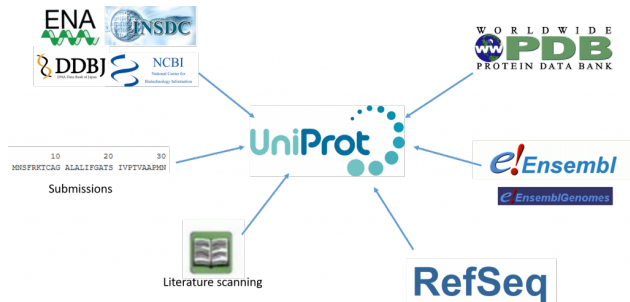
Curated families/ontologies

- Pfam
- KEGG
- EggNOG



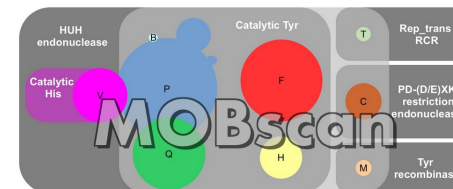
Large collections

- UniProt
- NCBI



Specialized databases

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



BUSCO



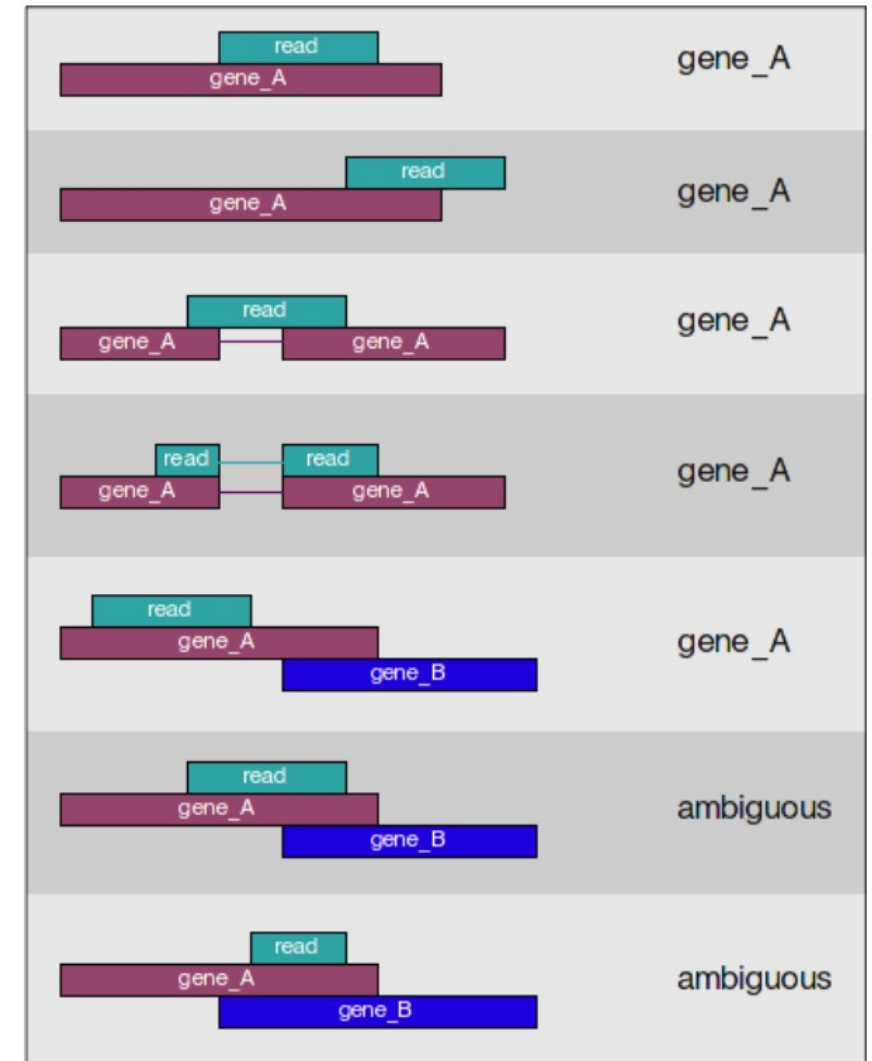
Counting reads



Reads per gene



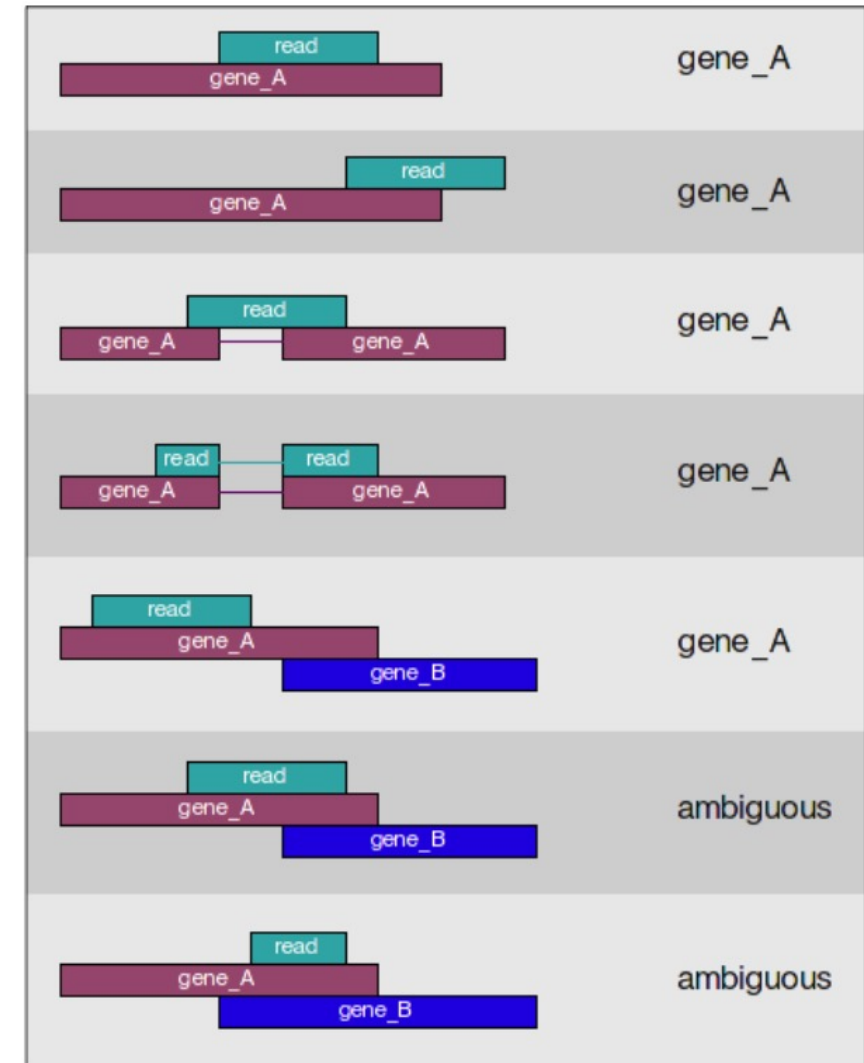
- e.g. featureCounts (subread) software
- count reads that map to positions overlapping with coding sequences
 - multi-overlap
 - length of overlap / overhang
 - stranding (direction)
 - forward- and reverse read



Reads per annotated functional class



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



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

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

- 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

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

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

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 $\log(\text{TPM}+1)$ 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](#)



Gordon Smyth ⚡ 45k
@gordon-smyth
Last seen 50 minutes ago
WEHI, Melbourne, Australia

4.8 years ago • updated 4.7 years ago **Gordon Smyth** ⚡ 45k

➤ no good statistics yet

<https://support.bioconductor.org/p/98820/#98875>

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_