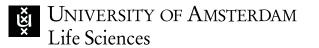
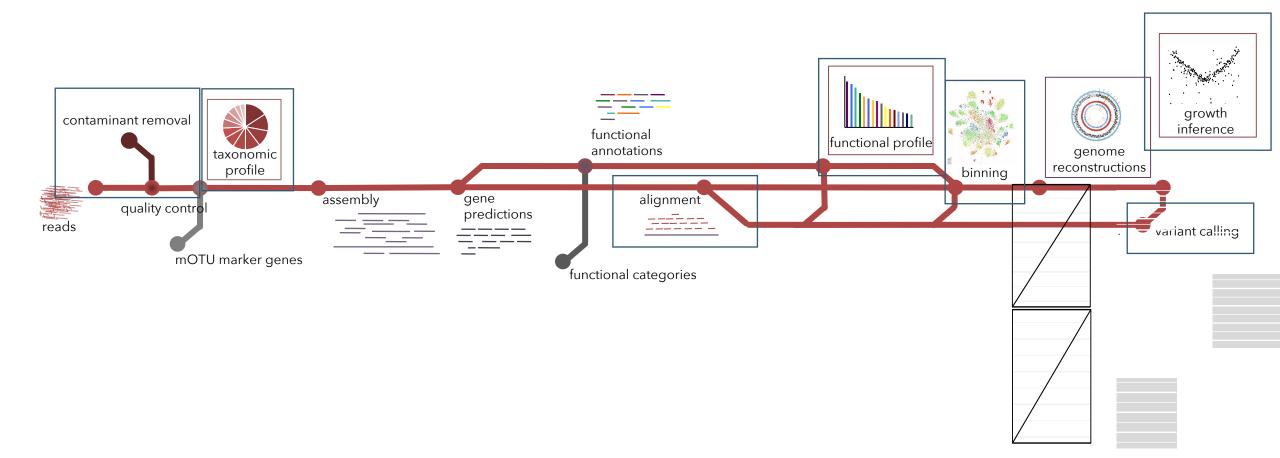
# Metagenomics 101

# Session 3: Read mapping 1

Anna Heintz-Buschart March 2022



### Metagenomics (+ other omics) pipeline

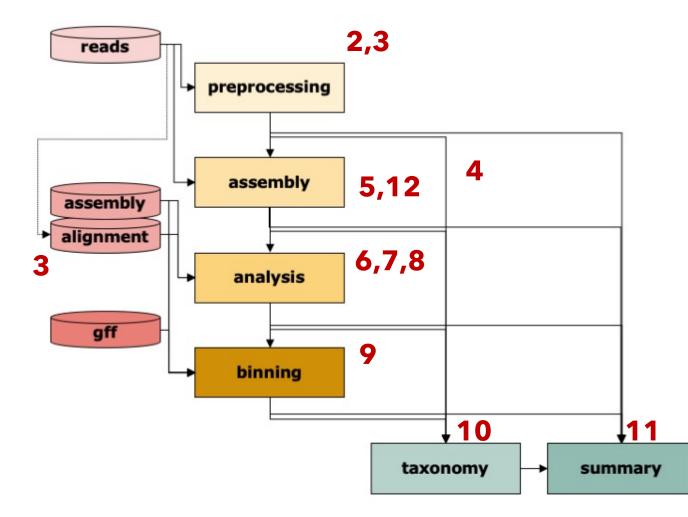


MP3



imp3.readthedocs.io

#### Metagenomics (+ other omics) pipeline

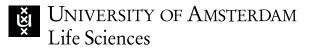


imp3.readthedocs.io



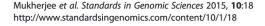
# Data preprocessing - remove contaminants!

- remove uninformative sequences:
- phiX spike-in
- host genome



#### Data preprocessing remove contaminants!

frontiers in Microbiology PERSPECTIVE published: 22 October 2021 doi: 10.3389/fmicb.2021.755101





**Open Access** 

#### COMMENTARY

#### Large-scale contamination of microbial isolate genomes by Illumina PhiX control

Supratim Mukherjee<sup>1\*</sup>, Marcel Huntemann<sup>1</sup>, Natalia Ivanova<sup>1</sup>, Nikos C Kyrpides<sup>1,2</sup> and Amrita Pati<sup>1</sup>

#### **Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics**

Valérian Lupo<sup>1,2</sup>, Mick Van Vlierberghe<sup>1</sup>, Hervé Vanderschuren<sup>3</sup>, Frédéric Kerff<sup>2</sup>, Denis Baurain<sup>1\*</sup> and Luc Cornet<sup>1,3</sup>

Steinegger and Salzberg *Genome Biology* (2020) 21:115 https://doi.org/10.1186/s13059-020-02023-1

Genome Biology



RESEARCH ARTICLE Removing contaminants from databases of draft genomes

Jennifer Lu<sup>1,2</sup>\*, Steven L. Salzberg<sup>1,2,3</sup>

#### METHOD

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

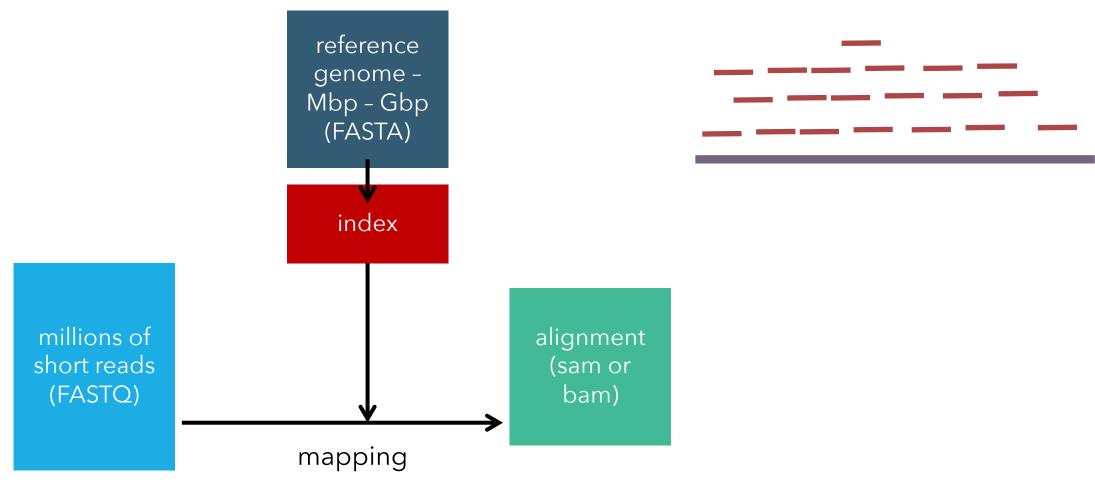
Martin Steinegger<sup>1,2,3\*</sup> and Steven L. Salzberg<sup>2,4,5</sup>





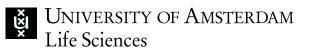


### Mapping reads

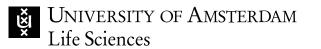


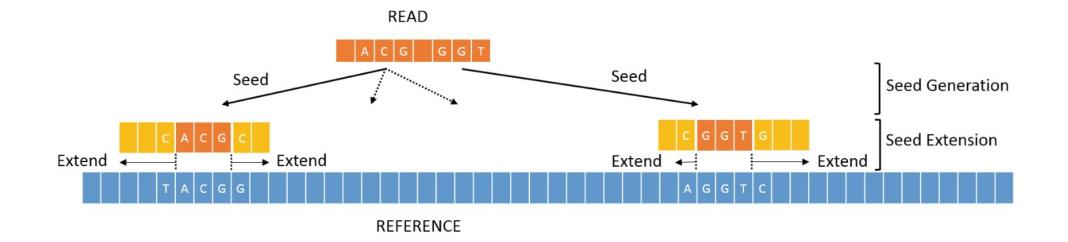
### Mapping reads





- the first alignment algorithms matched the query sequence (our read) to all the places in the reference (our selection / metagenome)
- dynamic programming was used to find the best match efficiently
- indexing is performed to enable the search for the places (seeds) where a read matches the reference
- dynamic programming is then used to expand the seed and retrieve the best match out of a few candidates







Houtgast et al. 2016. ARCS 2016

BWA-MEM:

#### BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 14 2009, pages 1754–1760 doi:10.1093/bioinformatics/btp324

Sequence analysis

## Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin\*



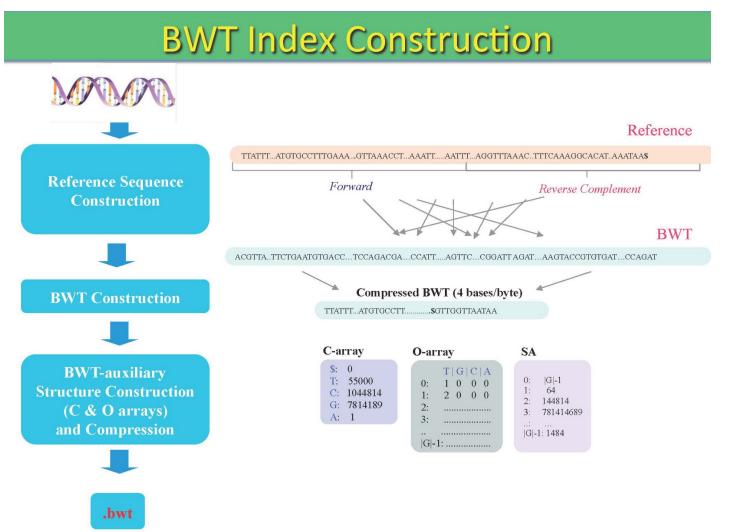


a. Seed extraction from reference genome **b**.Seed indexing using suffix tree or hash table characte location list 1<sup>st</sup> seed at 9 16 30 location 1 seed 60 7 5 12 3 reference genome 4 10 18 32 3 6 2 6 14 14 5 9 7 indexing seed location at the 16 60 12 reference genome location list 30 **C.**Seed extraction from reads **d**.Seed querying and filtering seeds 30 16 seed from from index data structure read 1: CCTTAGTATATACTAGTACGTT read 60 read 2: TATTCT FACGTAC FAGTACCGCCC 12 read 3: GCCTCTATATCCGTACTATATGGT reference genome e. Seed chaining and pre-alignment filtering **f**. Alignment verification CGTTAGTCTA . . . 0 0 0 0 0 read 1: CCTTAGTATATACTAGTACGTT 2 2 2 2 x 1 xx 1 X 5 5 6 6 ref 1: CGTTAGICTATATAATCO 9 9 9 9 9 bam/.sam file contains 5 7 9 11 11 11 11 11 necessary alignment 4 6 7 9 11 13 13 13 13 information (e.g., type, reference segment that spans 2 4 6 7 9 11 13 14 14 15 location, and number of locations (5, 7, and 9) each edit) 0 2 4 6 8 9 11 13 14 16 16

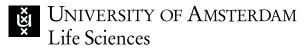


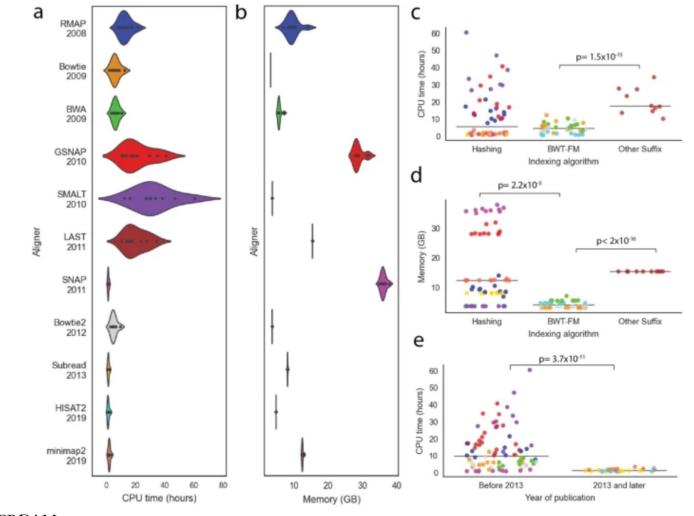
UNIVERSITY OF AMSTERDAM Life Sciences

Alser et al. 2021. Genome Biology 22: 249.

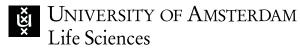






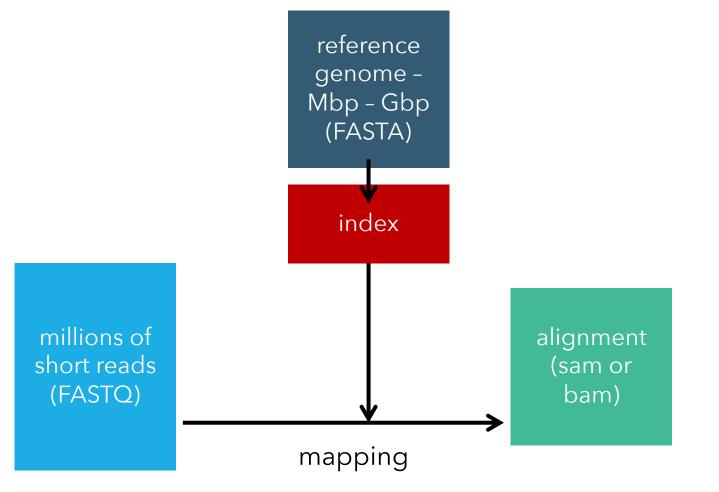






Alser et al. 2021. Genome Biology 22: 249.

### Mapping reads



#### Sacuanca alianments (SAM format)

JEQU			is (SAMINIOTTIAL)	
•	position	ame of read		
read name <u>flag</u>	reference quality p	a <u>rtn</u> er <u>~</u> a	alignment length read sequence	
SRR5947855.214477 99	test_contig_1 1 60 70M31 (CCFFDDFHHHGHGIJ.J0HI.GIIJJJ.IIJJJJGII	S = 1 69 I MERCEUN JJGHIIJJ MOS DUT		
:test_contig_94,14,-,38M63S,7 SRR5947855.214477 147 GAAGGAATCCAGAAGATATTCCAGGTG	.0; RG:Z:test test_contig_1 1 60 32S69 BDDDDDBDDCFFFFFHHFECFHGJIJIHEGJJJJJIJJ			distance to
:test_contig_2866,1838,+,66S3 SRR5947855.328367 99	5M,1,0; RG:Z:test test_contig_1 21 60 101M	= 136 216	TCCAAAGACAATGAAATAGCGAGAAGGAATCCAGGAAGATATTCCAGGTGCGATTTGCGGAATGATGACCTTTCT	reference
TAGTGCTTGGGCGGGCGTTGCATTTAA XS:i:0 RG:Z:test SRR5947855.326355 99	CCCFFFFFHFHHHJJJIJJJJJJJJGIIIIJJIIJJJF test_contig_1 24 60 101M	= 185 262		mismatching
TGCTTGGGCGGGCGTGCATTTAAATC XS:i:0 RG:Z:test SRR5947855.10987 99	CCCFFFFFHGHGHJJJJJJJJEBEIJJJJJJJEHGEGH test_contig_1 27 60 101M	IIGIJGIGJIIIIJIJIGFF>EHHF = 125 199		positions
TTGGGCGGGCGTTGCATTTAAATCTAA XS:i:0 RG:Z:test	CCCFFFFBFFHHGJJJIJIIJIIFJJJJJJGHIIJIIJJF	IJHHIIFHIJJJJIIIACCE>AHED	DFFFFFFECEED>>@BABDBBDDB@@ACDADCCDDDC> NM:i:0 MD:Z:101 AS:i:101	positions
SRR5947855.362726 163 TTTGGCTTTACGTTCAAATAAACGAAA XS:i:0 RG:Z:test	test_contig_1 91 60 101M CCCFFFFFHHHHHIJJJJJJJJIIIHHHHHFFFFFFEEE	= 141 151 DEEDDDDDEEEDDFEEEEDDDDDDE	TTCTTAGTGCTTGGGCGGGCGTTGCATTTAAATCTAATGCAGCTTCATATAATTCTGGATTZAIIIGIIGIAA) DEFEEEDDDDEDEEDDDDDDDDDDDDDDDDDDDDDDDD	
SRR5947855.10987 147 ATAAAACACAATGTCCTGCTAATAAAG	test_contig_1 125 60 101M DDCDDCDCCDDDFEDEEFFFEEHEHHHHHJJFIIIGEHI	= 27 -199 JJJIGHIHGB?DIIIGJIGCHHFHE	TAATGCAGCTTCATATAATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAG IEJJJIHEJIJIJIJIHFHHIJJJJIHF>HHFFFDFCC@NM:i:0MD:Z:101AS:i:101	alignment
XS:i:0 RG:Z:test SRR5947855.328367 147 TGTCCTGCTAATAAAGTCCAGTATGAG	test_contig_1 136 60 101M DDDEDDDDDDEEDEEEFFFFHHHHHHIGIJJJJGJJJ	= 21 -216 IIJJJJJJJJJJJJJJJJJJJJGCJJJ	CATATAATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAGATAAAACACAA JJJJJGHHJJJJJJJJJIIIJJJJJJJJHHGHHFFFFFCCC NM:i:0 MD:Z:101 AS:i:101	score
XS:i:0 RG:Z:test SRR5947855.362726 83 TGCTAATAAAGTCCAGTATGAGAAGAG	test_contig_1 141 60 101M DDDCDDDDEDEDECDDFFEFFHHHFHHJIHCIHHIJIIJ	= 91 -151	AATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAGATAAAACACAATGTCC GEIIJIIJIIJJIJIJIJIJIJIJIJIJHHHHHFFFFFCCC 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 XS:i:0 RG:Z:test SRR5947855.343465 99	DDDDDEEEEEFFFFFFFFFFGGIJIIIGDJJGJJJJJJ test_contig_1 204 60 101M	JJJIJJJJJJJIHJJJJJJJJJJJJIJIHF = 257 154	IFJIJIGJIGIIIJIHGIJJIGIJJIHFGHHFFFDD@B@ NM:i:0 MD:Z:101 AS:i:101 ACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAG	
	test_contig_1 204 60 101M	= 257 154	ACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAG	

CATAACTGTCAATGATAAAGCCATGAC NM:i:0 MD:Z:101 AS:i:101 XS:i:0 RG:Z:test 99 test\_contig\_1 257 60 101M 290 134 GATATTACCTGCAAAAACAAACATAACTGTCAATGATAAAGCCATGACGACTTCAGCATTTACAACAGGAATTT NM:i:0 MD:Z:101 AS:i:101 GAGTCATTCCTTCGATAACAGTTTGAG XS:i:0 RG:Z:test

read quality

SRR5947855.343465 147 CTCAATCATA AAGCCATGACGACTTCAGCATTTACAACAGGAATTT GAGTCATTCCTTCGATAACAGTTTGAG NM:i:0 MD:Z:101 AS:i:101 XS:i:0 RG:Z:test

CIGAR position of partner



SRR5947855.285811

×X×

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	$\operatorname{Int}$	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	\* [:rname:^*=][:rname:]*	Reference sequence NAME <sup>11</sup>
4	POS	$\operatorname{Int}$	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^8-1]$	MAPping Quality
6	CIGAR	String	\* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* = [:rname:^*=][:rname:]*	Reference name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	\* [A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33



I	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



Ē	Bit	Description	-	
1	0x1	template having multiple segments in sequencing	read1	
2	0x2	each segment properly aligned according to the aligner	IEaui	
4	0x4	segment unmapped refere		
8	0x8	next segment in the template unmapped		-
16	0x10	SEQ being reverse complemented		read2
32	0x20	SEQ of the next segment in the template being reverse complemented		TEAUZ
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	_	



Е	Bit	Description				
1	0x1	template having multiple segments in sequencing				
2	0x2	each segment properly aligned according to the aligner		read	read2	
4	0x4	segment unmapped	referer			
8	0x8	next segment in the template unmapped	lelelel			
16	0x10	SEQ being reverse complemented				
32	0x20	SEQ of the next segment in the template being reverse complete	mented			
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				



E	Bit	Description	-
1	0x1	template having multiple segments in sequencing	
2	0x2	each segment properly aligned according to the aligner	read1
4	0x4	segment unmapped refer	ence
8	0x8	next segment in the template unmapped	ence
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	



E	Bit	Description	
1	0x1	template having multiple segments in sequencing	
2	0x2	each segment properly aligned according to the aligner	
4	0x4	segment unmapped	reference
8	0x8	next segment in the template unmapped	IEIEIEIIC
16	0x10	SEQ being reverse complemented	
32	0x20	SEQ of the next segment in the template being reverse con	nplemented
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 control	s
1024	0x400	PCR or optical duplicate	
2048	0x800	supplementary alignment	





В	it	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



B	it	Description
1	0x1	template having multiple segments in sequencing
$rac{2}{4}$	0x2	each segment properly aligned according to the aligner
<b>4</b>	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32 04	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
1000		



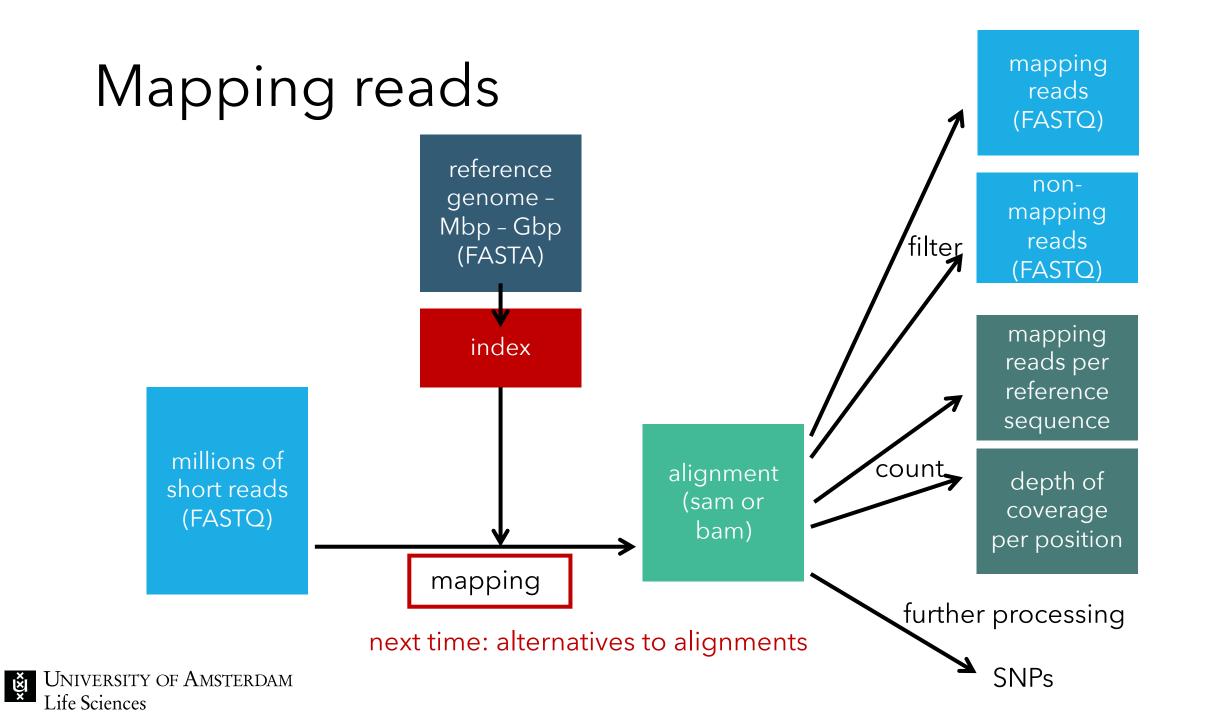
B	$\mathbf{it}$	Description
1	0x1	template having multiple segments in sequencing
$\frac{2}{4}$	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
$\frac{16}{32}$	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment 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



#### flag

SRR5947855.214477 99 TCTTATACACATCTCCGAGCCCACGAG	test_contig_1 1 60 70M31S = 1 69 AAGCAGTGACTATGTAGTCATCCAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCTGTC @CCFFDDFHHHGHGIJJJQHIJGIIJJJJIJJJJGIIIJ6F0DFHGJJGHIIJJJG>DHIJJDA@DHBEHHHGEHBDDDDEDEEEECDDDBDDDDDBD NM:i:0 MD:Z:70 AS:i:70 XS:i:0 SA:Z
SRR5947855.214477 147	test_contig 1 1 60 32S69M = 1 -69 CGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTGACTATGTAGTCATCCAAAGACAATGAAATAGCGA
GAAGGAATCCAGAAGATATTCCAGGTG	BDDDDDBDDCFFFFHHFECFHGJIJIHEGJJJJJIJGGEIJJHGIIJJJIIHIJIJIHGIHBIJJJJGHEDCIDHGJJHHBHFFFFF@@@ NM:i:0 MD:Z:69 AS:i:69 XS:i:0 SA:Z
:test_contig_2866,1838,+,66S3	
SRR5947855.328367 99	test_contig_1 21 60 101M = 136 216 TCCAAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGATTTGCGGAATGATCATCTTCT
TAGTGCTTGGGCGGGCGTTGCATTTAA	CCCFFFFFHHHHJJJJJJJJJJJJJJJJJJJJJIJIJJJJJJJJ
XS:i:0 RG:Z:test	
SRR5947855.326355 99	test_contig_1 24 60 101M = 185 262 AAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGATTTGCGGAATGATGATGACCTTTCTTAG
TGCTTGGGCGGGCGTTGCATTTAAATC	CCCFFFFFHGHGHJJJJJJJJEBEIJJJJJJJEHGEGHIIGIJGIGJIIIIJJJIGFF>EHHFFFFEEEEDDCCCDDDDDDDDDDDDDC9>BDDDEDEACDC NM:i:0 MD:Z:101 AS:i:101
XS:i:0 RG:Z:test	
SRR5947855.10987 99	test_contig_1 27 60 101M = 125 199 GACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGAATTGCGGAATGATGACCTTTCTTAGTGC
TTGGGCGGGCGTTGCATTTAAATCTAA	CCCFFFFBFFHHGJJJIJIIJIIFJJJJJJGHIIJIIJFIJHHIIFHIJJJJIIIACCE>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 60 101M = 141 151 TTCTTAGTGCTTGGGCGGGCGTTGCATTTAAATCTAATGCAGCTTCATATAATTCTGGATTCATTGTGTAAC
TTTGGCTTTACGTTCAAATAAACGAAA	CCCFFFFFHHHHHIJJJJJJJJIIIHHHHHFFFFFFEEEDEEDDDDDEEEDDFEEEEDDDDDEFEEEDDDDDD
XS:i:0 RG:Z:test	
SRR5947855.10987 147	test_contig_1 125 60 101M = 27 -199 TAATGCAGCTTCATATAATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAG
ATAAAACACAATGTCCTGCTAATAAAG	DDCDDCDCCDDDFEDEEFFFEEHEHHHHHHJJFIIIGEHIJJJIGHIHGB?DIIIGJIGCHHFHEJJJIHEJIJIJIJIHFHHIJJJJIHF>HHFFDFCC@NM:i:0MD:Z:101AS::101AS::101
XS:i:0 RG:Z:test	
SRR5947855.328367 147	test_contig_1 136 60 101M = 21 -216 CATATAATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAGATAAAACACAA
TGTCCTGCTAATAAAGTCCAGTATGAG	DDDEDDDDDDDEEDEEEFFFFFHHHHHHHIGIJJJJGJJJIJJJJJJJJJJJJJJJJ
XS:i:0 RG:Z:test	
SRR5947855.362726 83	test_contig_1 141 60 101M = 91 -151 AATTCTGGATTCATTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAGATAAAACACAAATGTCC
TGCTAATAAAGTCCAGTATGAGAAGAG	DDDCDDDDEDEDECDDFFEFFHHHFHHJIHCIHHIJIIJJJJJJJJJJJJJJJJJJJJJ
XS:i:0 RG:Z:test	test contig 1 185 60 101M = 24 -262 AACGAAAGGCATAGATAAAACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAG
SRR5947855.326355 147 TATTACCTGCAAAAACAAACATAACTG	test_contig_1 185 60 101M = 24 -262 AACGAAAGGCATAGATAAAACACAAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAG
XS:i:0 RG:Z:test	
SRR5947855.343465 99	test_contig_1 204 60 101M = 257 154 ACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAG
CATAACTGTCAATGATAAAGCCATGAC	CCCFFFFFHHGHHJJJJJJJJJJHHIJJHHIIGEEGFGGGHIIJIIJJJFIJIIFIJJJIJIJJJJJIJJJJJJJJJJ
XS:i:0 RG:Z:test	
SRR5947855.285811 99	test contig 1 257 60 101M = 290 134 GATATTACCTGCAAAAACAAACATAACTGTCAATGATAAAGCCATGACGACTTCAGCATTTACAACAGGAATTT
GAGTCATTCCTTCGATAACAGTTTGAG	CCCFFFFFHHGHHCEGBHIIIIAHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
XS:i:0 RG:Z:test	
SRR5947855.343465 147	test_contig_1 257 60 101M = 204 -154 GATATTACCTGCAAAAACAAACATAACTGTCAATGATAAAGCCATGACGACTTCAGCATTTACAACAGGAATTT
GAGTCATTCCTTCGATAACAGTTTGAG	DFDCCC?DDEDDDEEFFFFFHHHHHHHIJJJJJJJJJJJJJJJJJJIJIIIIIJIHHGIIIECJJJJJJJJJJJJJJJJJJJHHBJJFJJJHIHHHGHFFFFCC@ NM:i:0 MD:Z:101 AS:i:101
XS:i:0 RG:Z:test	

×××





#### Thanks for your attention!



a.u.s.heintzbuschart@uva.nl

SP C2.205



github.com/a-h-b



twitter.com/\_a\_h\_b\_

