Metagenomics 101

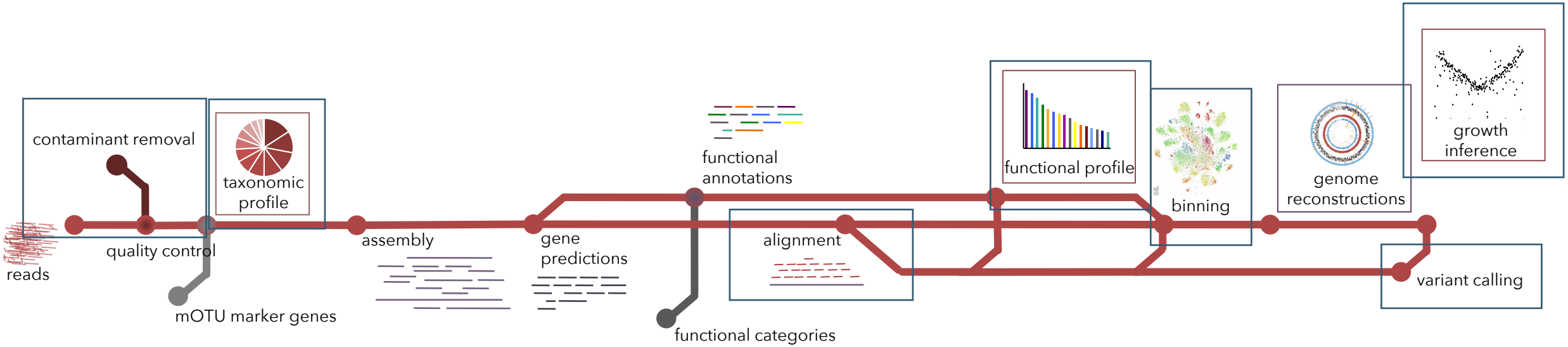
Session 3: Read mapping 1

Anna Heintz-Buschart

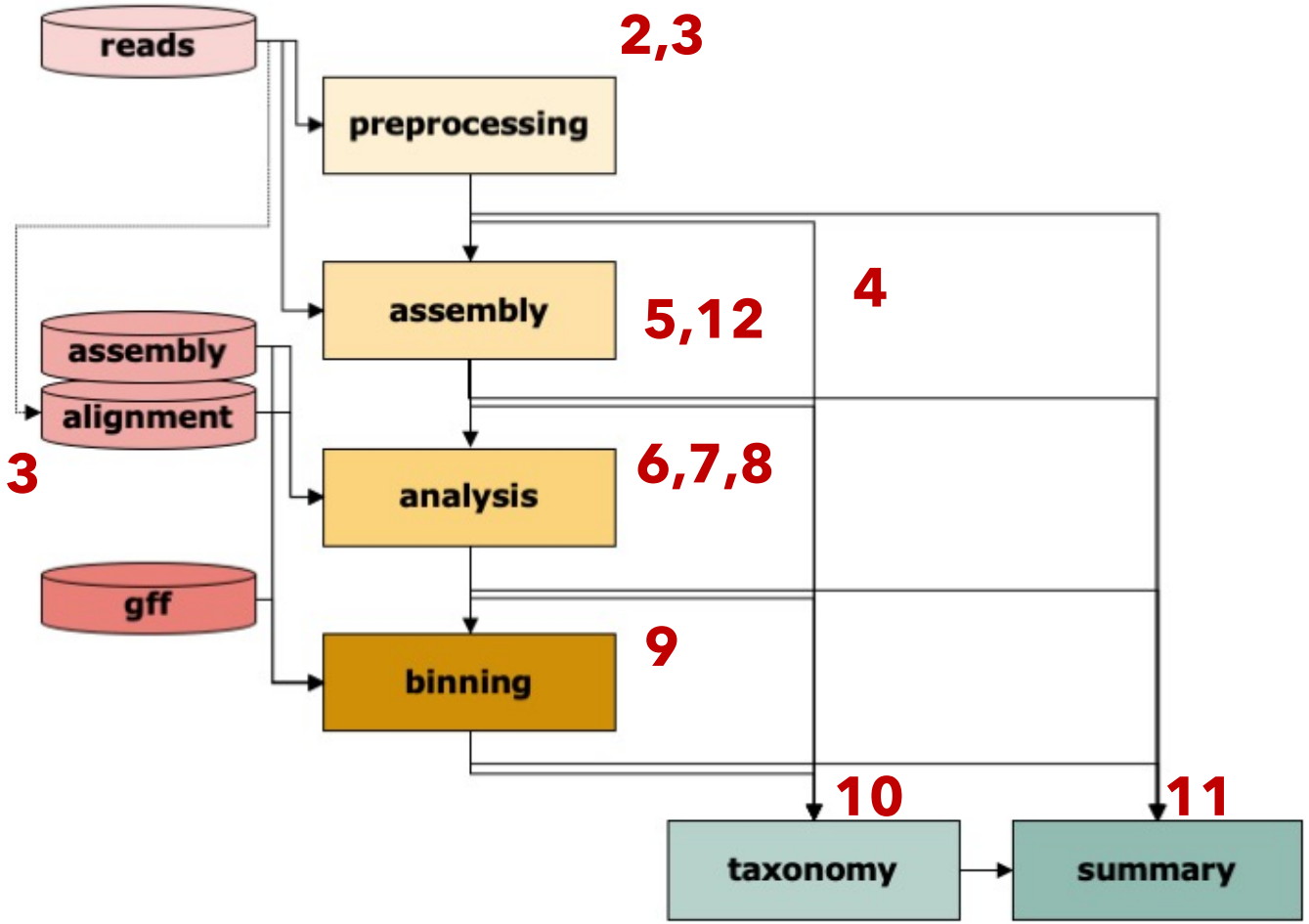
March 2022

University of Amsterdam
Life Sciences

Metagenomics (+ other omics) pipeline



Metagenomics (+ other omics) pipeline



Data preprocessing – remove contaminants!

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

Data preprocessing – remove contaminants!

Mukherjee *et al. Standards in Genomic Sciences* 2015, **10**:18
http://www.standardsingenomics.com/content/10/1/18

Standards in Genomic Sciences

COMMENTARY Open Access

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

Supratim Mukherjee[1*], Marcel Huntemann[1], Natalia Ivanova[1], Nikos C Kyrpides[1,2] and Amrita Pati[1]

frontiers in Microbiology

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

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

Valérian Lupo[1,2], Mick Van Vlierberghe[1], Hervé Vanderschuren[3], Frédéric Kerff[2], Denis Baurain[1] and Luc Cornet[1,3]*

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

Genome Biology

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Removing contaminants from databases of draft genomes

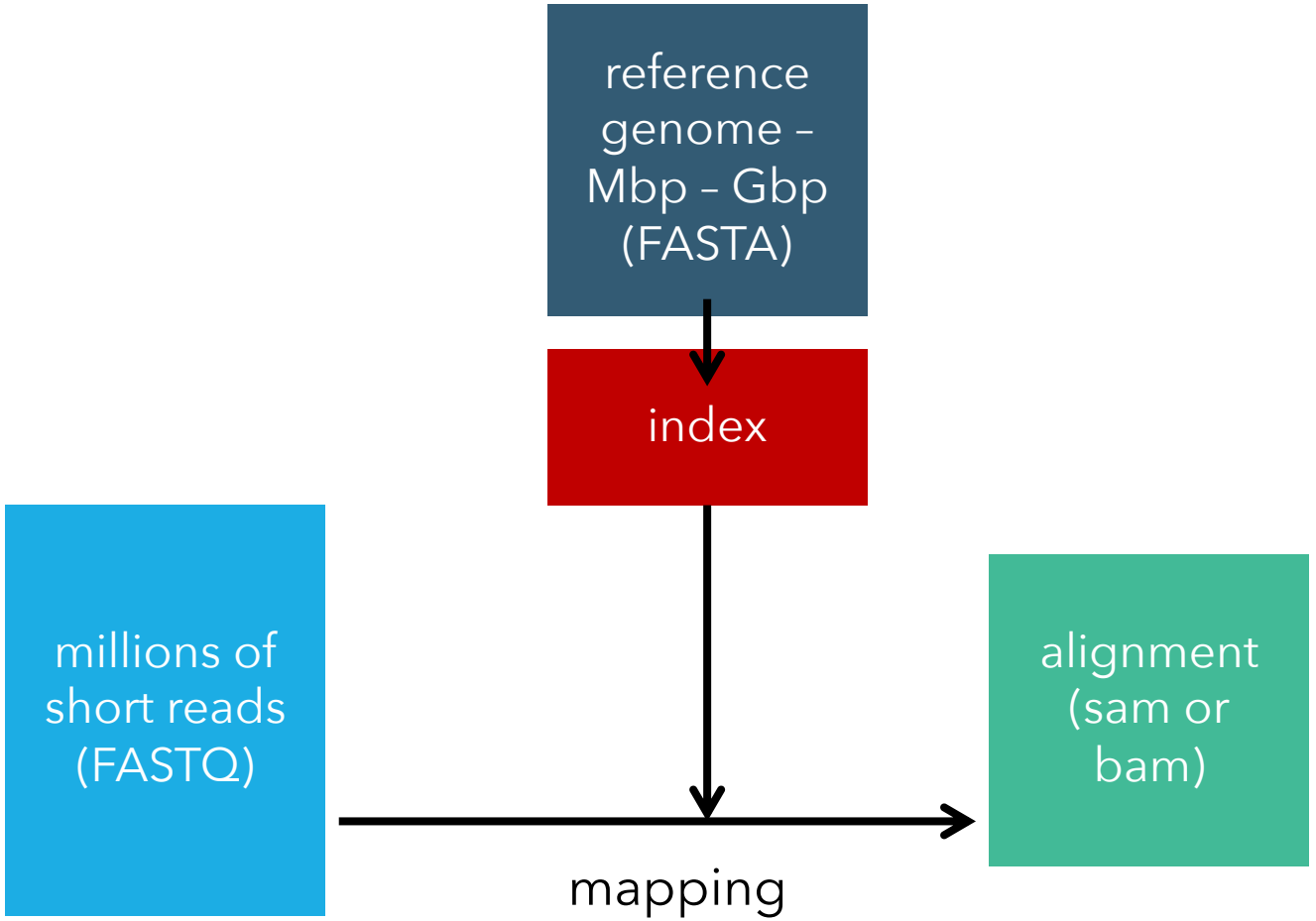
Jennifer Lu[1,2]*, Steven L. Salzberg[1,2,3]

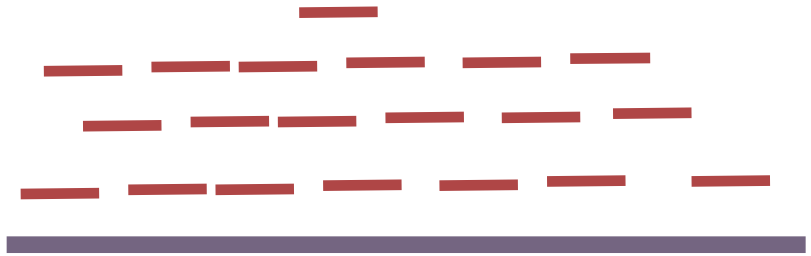
METHOD Open Access

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

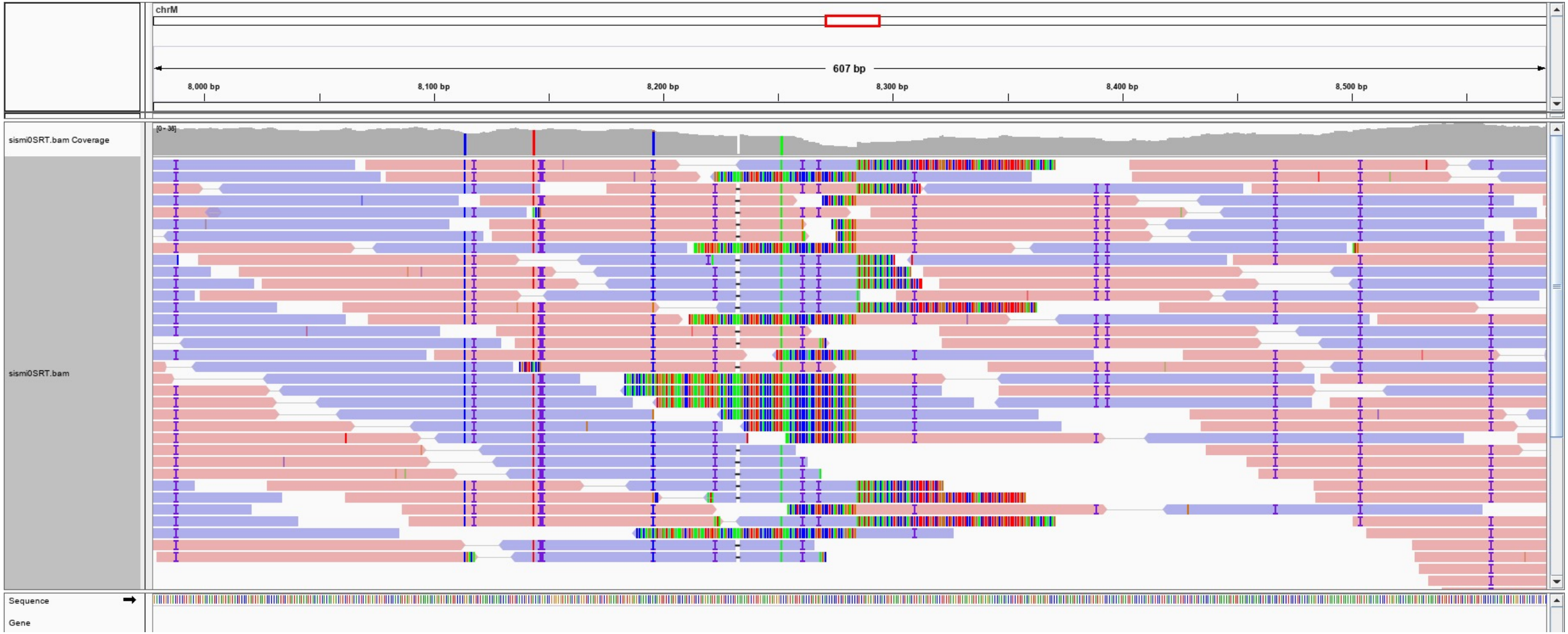
Martin Steinegger[1,2,3*] and Steven L. Salzberg[2,4,5]

Mapping reads





Mapping reads



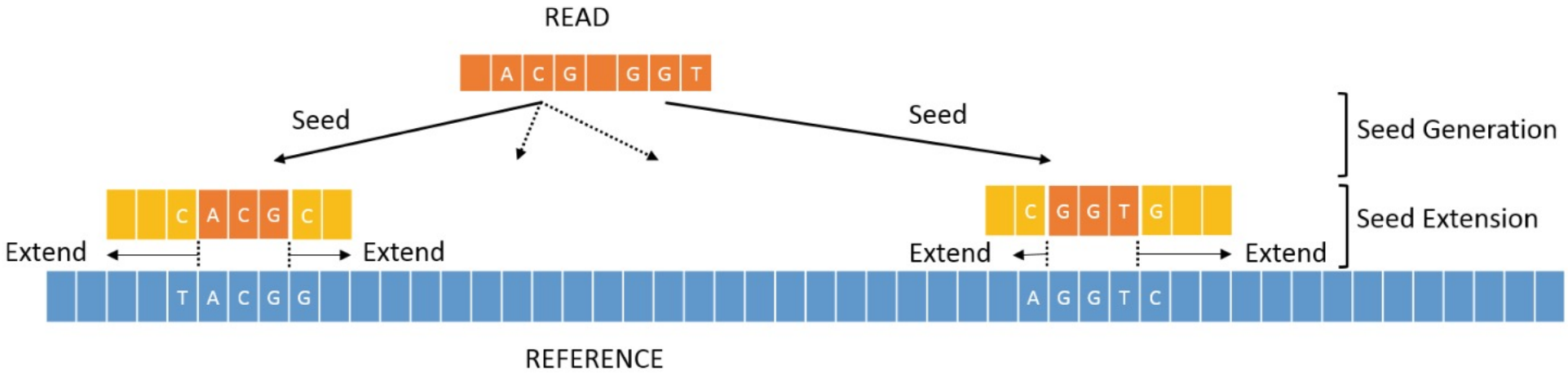
Mapping reads = aligning reads

- the first alignment algorithms matched the query sequence (our read) to all the places in the reference (our genome / metagenome)
- dynamic programming was used to find the best match efficiently

too slow

- 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

Mapping reads = aligning reads



Houtgast et al. 2016. ARCS 2016

Mapping reads = aligning reads

BWA-MEM:

BIOINFORMATICS ORIGINAL PAPER

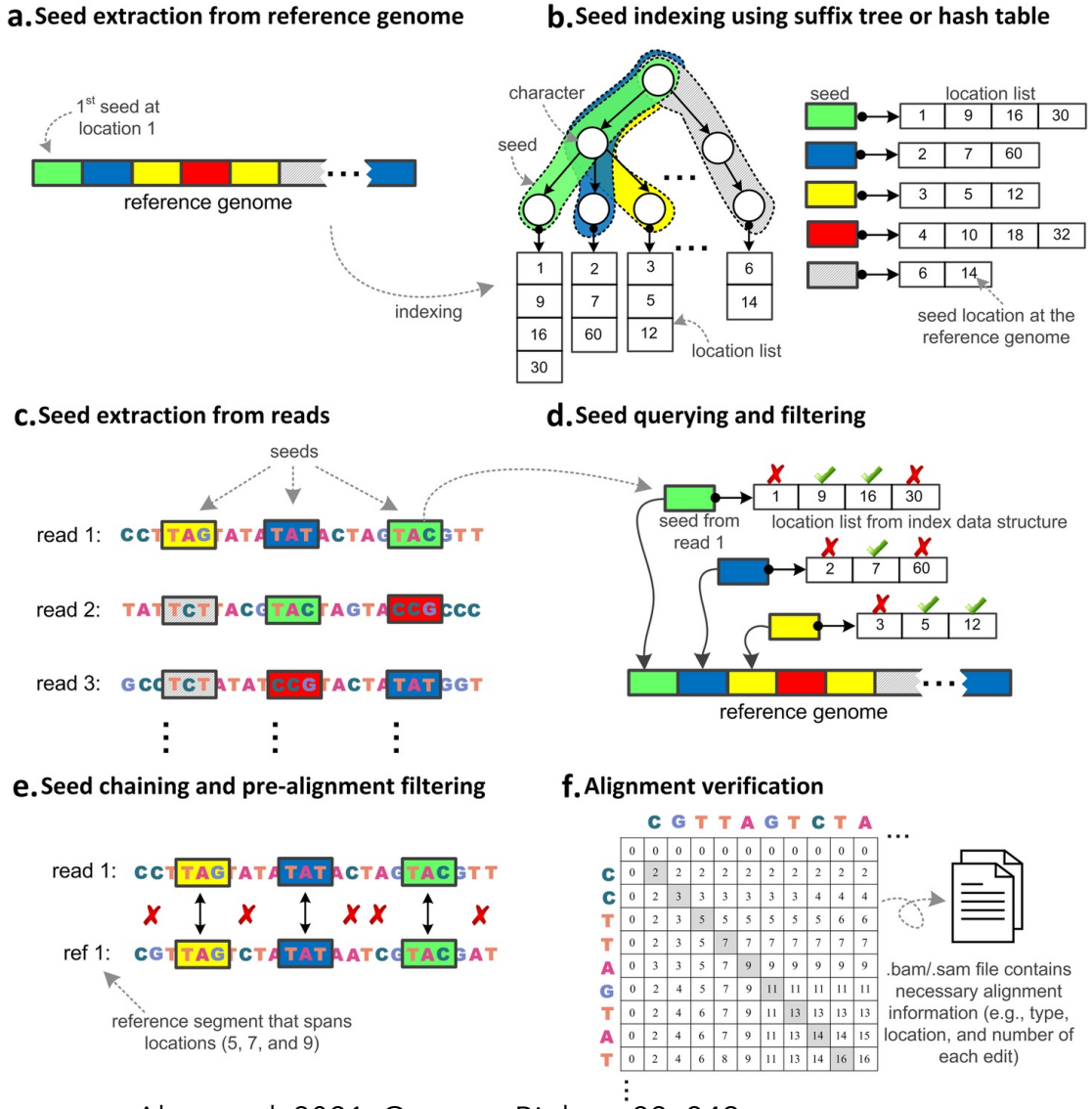
Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

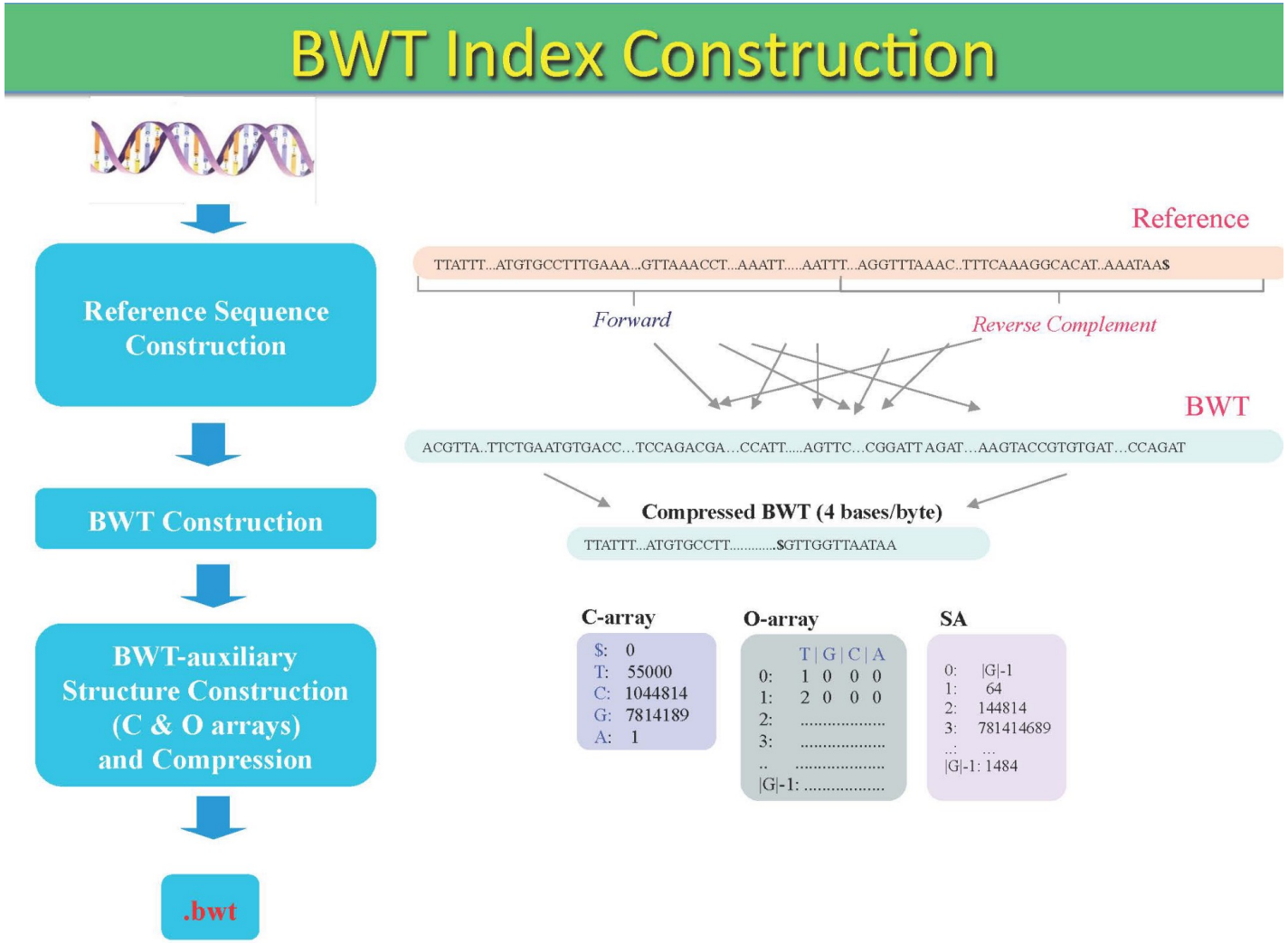
Heng Li and Richard Durbin*

Mapping reads = aligning reads

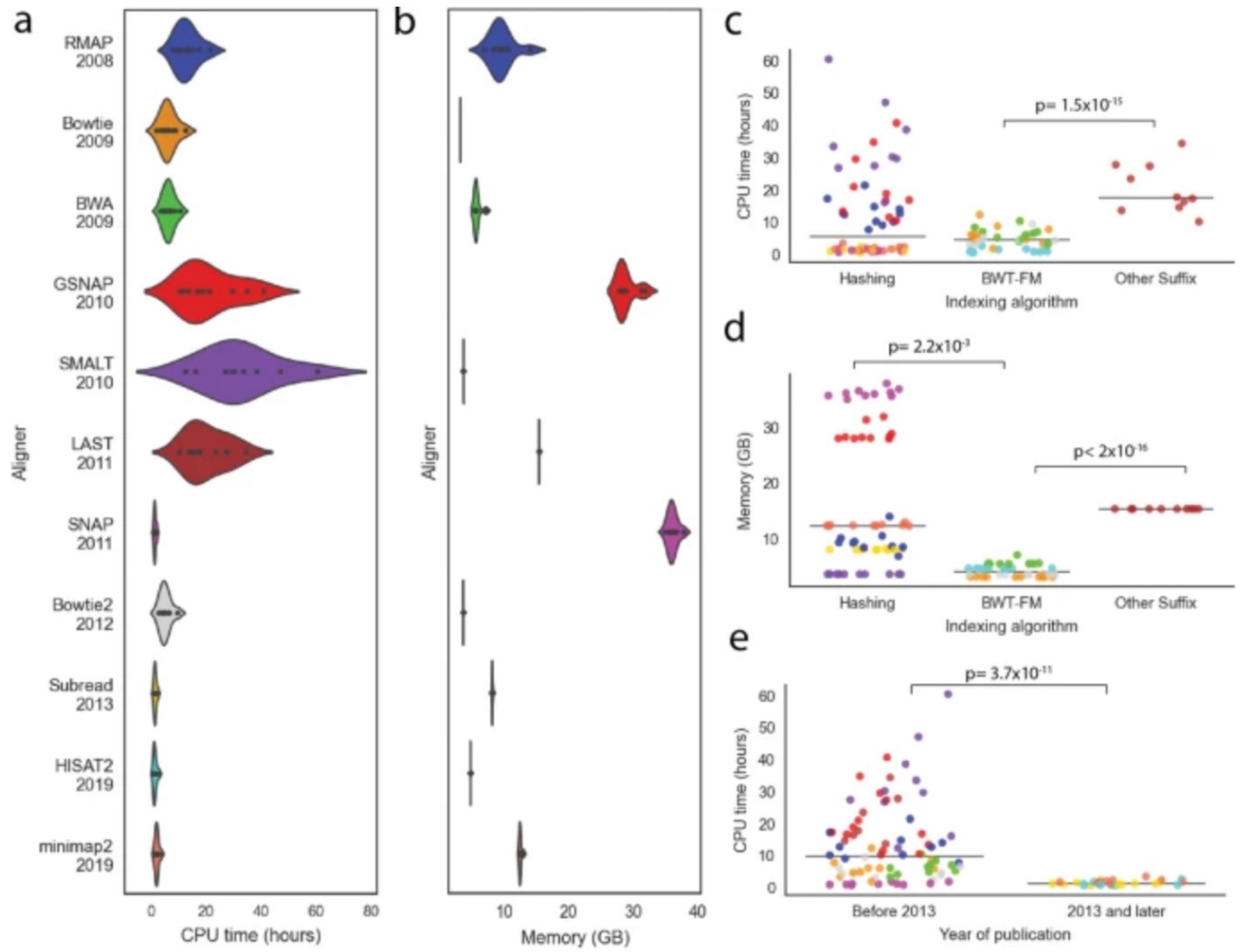


Alser et al. 2021. Genome Biology 22: 249.

Mapping reads = aligning reads

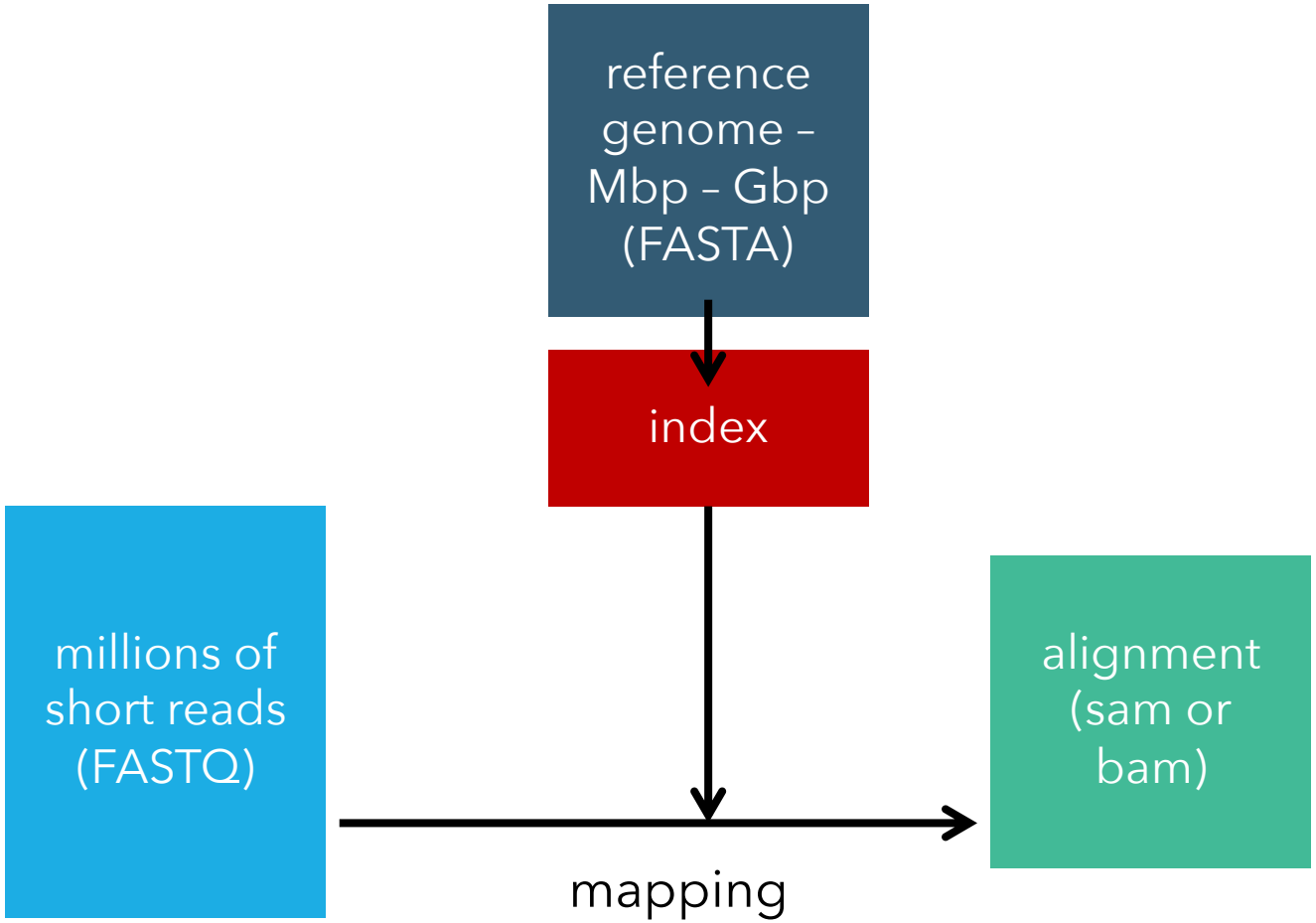


Mapping reads = aligning reads



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

Mapping reads



Sequence alignments (SAM format)

Sequence alignments (SAM format)

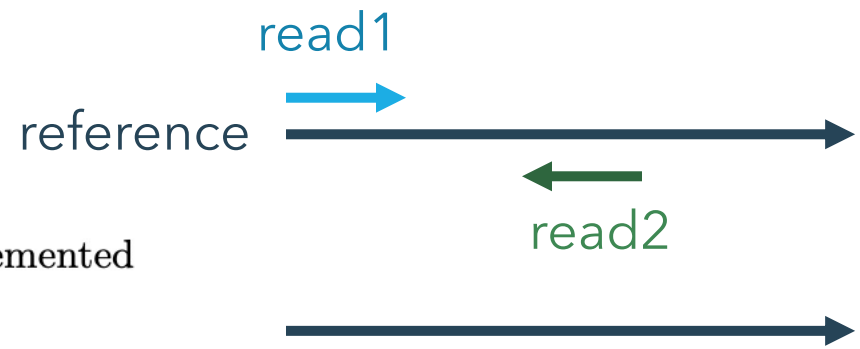
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	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

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

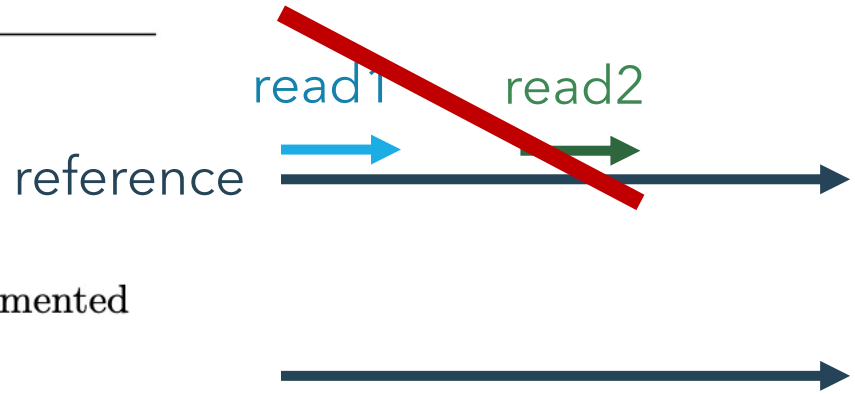
Sequence alignments (SAM format)

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



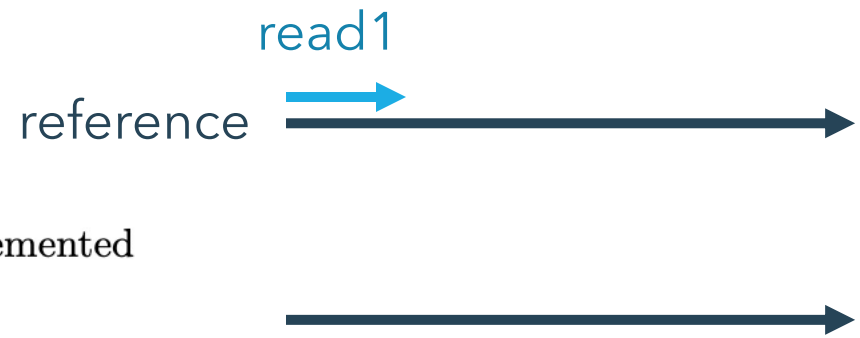
Sequence alignments (SAM format)

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



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



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

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

83

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

99

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

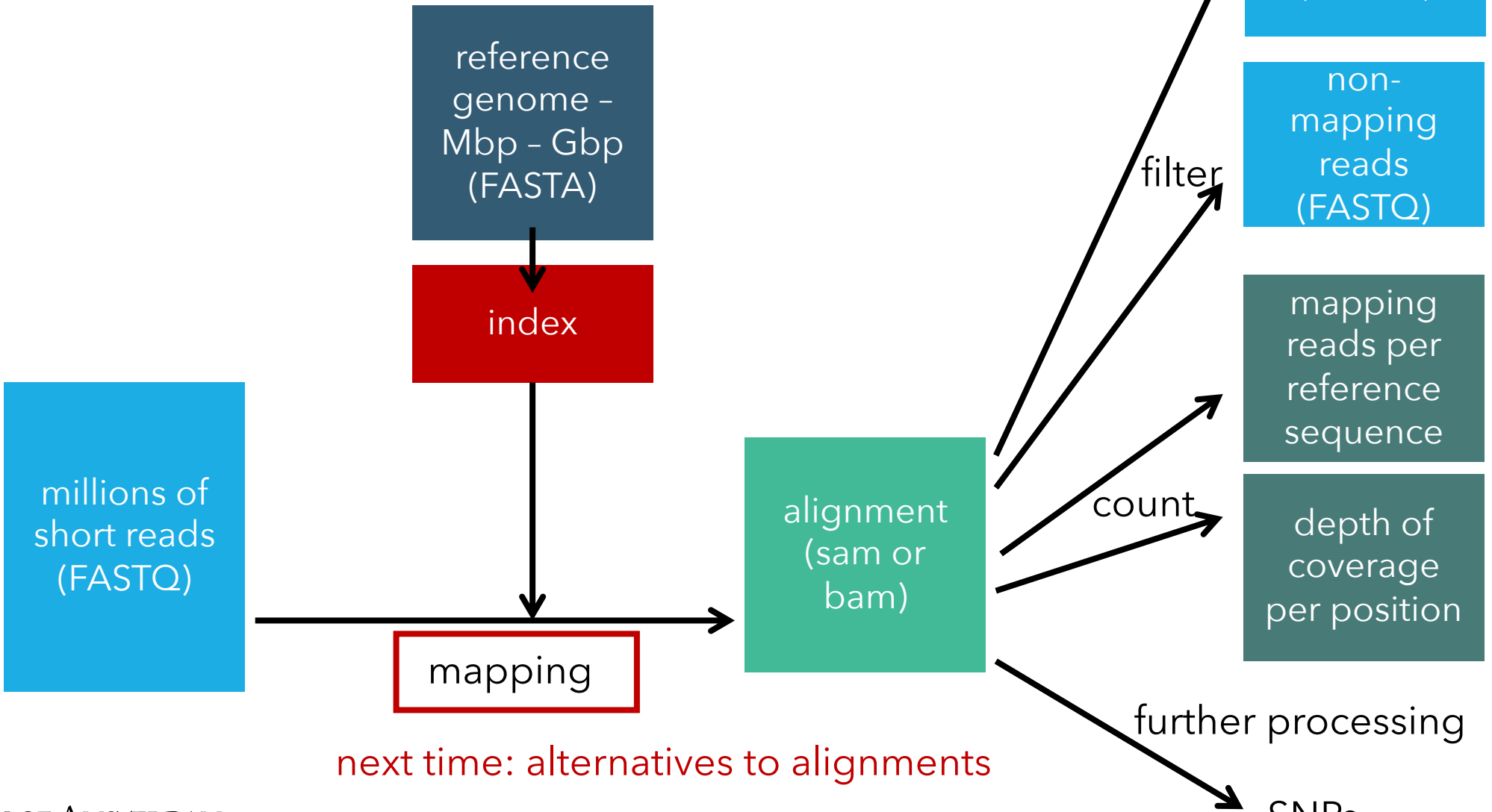
147

Sequence alignments (SAM format)

flag

```
SRR5947855.214477       99      test_contig_1   1       60      70M31S  =       1       69      AAGCAGTGACTATGTAGTCATCCAAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCTGTC
TCTTATACACATCTCCGAGCCCACGAG   @CCFFDDFHHHGHGIJJIJ@HIJGIIJJJJIIJJJJGIIIJ6F0DFHGJJGHIIJJJIG>DHIJJDA@DHBEHHHGEHBDDDDEDEEEECDDDBDDDDDBD   NM:i:0  MD:Z:70 AS:i:70 XS:i:0  SA:Z
:test_contig_94,14,-,38M33S,7,0;    RG:Z:test
SRR5947855.214477       147     test_contig_1   1       60      32S69M  =       1       -69     CGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGCAGTGACTATGTAGTCATCCAAAGACAATGAAATAGCGA
GAAGGAATCCAGAAGATATTCCAGGTG   BDDDDDBDDCFFFFFHHFECFHGJIJIHEGJJJJJIJJJJIIGGEIJJIHGIIJJJIIIHIIJIIIHGIIHBIJJJIGHEDCIDHGJJHHBHFFFFFF@@@   NM:i:0  MD:Z:69 AS:i:69 XS:i:0  SA:Z
:test_contig_2866,1838,+,66S35M,1,0;        RG:Z:test
SRR5947855.328367       99      test_contig_1   21      60      101M    =       136     216     TCCAAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGATTTGCGGAATGATGACCTTTCT
TAGTGCTTGGGCGGGCGTTGCATTTAA   CCCFFFFFHFHHHJJJIJJIJJJJJJGIIIIJJIIJJJFIJIJJJJIHHIJJJJJJIJIGHHFFFFFEEEEEEDDDCDDEDDDDDDDDDDDDDDDDEDEA   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
SRR5947855.326355       99      test_contig_1   24      60      101M    =       185     262     AAAGACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGATTTGCGGAATGATGACCTTTCTTAG
TGCTTGGGCGGGCGTTGCATTTAAATC   CCCFFFFFHGHGHJJJJJJJJJEBEIJJJJJJJJJEHGEGHIIGIJGIGJIIIIJIJIGFF>EHHFFFFFEEEEDDCCCDDDDDDDDDDC9>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     GACAATGAAATAGCGAGAAGGAATCCAGAAGATATTCCAGGTGCGATTTGCGGAATGATGACCTTTCTTAGTGC
TTGGGCGGGCGTTGCATTTAAATCTAA   CCCFFFFBFFHHGJJJIJIIJIIFJJJJJGHIIJIIJJFIJHHIIFHIJJJJIIIACCE>AHEDFFFFFFECEED>>@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     TTCTTAGTGCTTGGGCGGGCGTTGCATTTAAATCTAATGCAGCTTCATATAATTCTGGATTCATTTGTTGTAAC
TTTGGCTTTACGTTCAAATAAACGAAA   CCCFFFFFFHHHHHIJJJJJJJIIIHHHHHHFFFFFFFEEEDEEDDDDDEEEDDFEEEEDDDDDDEFEEEDDDDEDEEDDDDDDDDDDDDDDEDDDDDDDDD   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
SRR5947855.10987        147     test_contig_1   125     60      101M    =       27      -199    TAATGCAGCTTCATATAATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAG
ATAAAACACAATGTCCTGCTAATAAAG   DDCDDCDCCDDDFEDEEFFFEEHEHHHHHHJJFIIIGEHIJJJIGHIHGB?DIIIGJIGCHHFHEJJJIHEJIJIJIJIHFHHIJJJJIHF>HHFFFDFCC@   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
SRR5947855.328367       147     test_contig_1   136     60      101M    =       21      -216    CATATAATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAGATAAAACACAA
TGTCCTGCTAATAAAGTCCAGTATGAG   DDDEDDDDDDDEEDEEEFFFFFHHHHHHHIGIJJJJGJJJIIJJJJJJJJIJJIJJJJJIGCJJJIJJGHHJJJJJJJIIIJJJJJJJJHHGHHFFFFFCCC   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
SRR5947855.362726       83      test_contig_1   141     60      101M    =       91      -151    AATTCTGGATTCATTTGTTGTAACTTTGGCTTTACGTTCAAATAAACGAAAGGCATAGATAAAACACAATGTCC
TGCTAATAAAGTCCAGTATGAGAAGAG   DDDCDDDDEDEDECDDFFEFFHHHFHHJIHCIHHIJIIJJJJJJJJGJJIJJJJJJJJJJJJIIGEIIJIIIJIIIJJIJIJIIIJIIIJHHHHHFFFFFCCC   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    AACGAAAGGCATAGATAAAACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAATTTCTAAATATGA
TATTACCTGCAAAAACAAACATAACTG   DDDDDEEEEEFFFFFFFFEFHGGIJIIIGDJJGJJJJJJJJJJJIJJJJJJIHJJJJJIJJIJIHFJIJIGJIGIIIJIHGIJJIGIJJIHFGHHFFFDD@B@   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
SRR5947855.343465       99      test_contig_1   204     60      101M    =       257     154     ACACAATGTCCTGCTAATAAAGTCCAGTATGAGAAGAGAGAATTTCTAAATATGATATTACCTGCAAAAACAAA
CATAACTGTCAATGATAAAGCCATGAC   CCCFFFFFHHGHHJJJJJJJJJJHHIJJHHIIGEEGFGGGHIIJIIJJJFIJIIFIJJJIJIJIIIJJJJIIJJIJJDGFHHGGEFFFFFFFEEEE=CDDDDD   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
SRR5947855.285811       99      test_contig_1   257     60      101M    =       290     134     GATATTACCTGCAAAAACAAACATAACTGTCAATGATAAAGCCATGACGACTTCAGCATTTACAACAGGAATTT
GAGTCATTCCTTCGATAACAGTTTGAG   CCCFFFFFHHGHHCEGBHIIIIAHIIIIIIIIIIIIIIIIIIIHICIIIIGIIIGIIIIIIIGIICHHIIHFHGHGDDBEFFEEFDEEDDDD@DDDCDEDDD   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
SRR5947855.343465       147     test_contig_1   257     60      101M    =       204     -154    GATATTACCTGCAAAAACAAACATAACTGTCAATGATAAAGCCATGACGACTTCAGCATTTACAACAGGAATTT
GAGTCATTCCTTCGATAACAGTTTGAG   DFDCCC?DDEDDDEEFFFFFHHHHHHHIIJIJJIJJJJJJJJJJIIJIIIIIJIHHGIIIIECJJJJJJJJJIJJJHHBJJIFJJJHIIHHHGHFFFFFCC@   NM:i:0  MD:Z:101        AS:i:101
        XS:i:0  RG:Z:test
```

Mapping reads



next time: alternatives to alignments

Thanks for your attention!

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_

University of Amsterdam
Life Sciences