

ISO/IEC 23092-2:2019-10 (E)

Information technology - Genomic information representation - Part 2: Coding of genomic information

Contents		Page
Foreword		vi
Introduction		vii
1	Scope	1
2	Normative references	1
3	Terms and definitions	1
4	Abbreviations	6
5	Conventions	6
5.1	General	6
5.2	Arithmetic operators	6
5.3	Logical operators	7
5.4	Relational operators	7
5.5	Bit-wise operators	7
5.6	Assignment operators	8
5.7	Range notation	8
5.8	Mathematical functions	8
5.9	Order of operation precedence	9
5.10	Variables, syntax elements and tables	10
5.11	Text description of logical operators	11
5.12	Processes	12
6	Syntax and semantics	12
6.1	Method of specifying syntax in tabular form	12
6.2	Bit ordering	13
6.3	Specification of syntax functions and data types	13
6.4	Semantics	14
7	Data structures	15
7.1	Data unit	15
7.2	Raw reference	16
7.2.1	Syntax and semantics	16
7.3	Parameter set	16
7.3.1	Syntax and semantics	16
7.3.2	Encoding parameters	17
7.4	Access unit	23
7.4.1	Syntax and semantics	23
7.4.2	Access unit types	27
8	Descriptors	27
9	Sequencing reads	30
9.1	Supported symbols	31
9.2	Paired-end reads	32
9.3	Reverse-complement reads	32
9.4	Data classes	33
9.5	Aligned data	33

9.6	Unaligned data	34
10	Decoding process	35
10.1	General	35
10.2	dataset_type = 0 or 1	35
10.2.1	References padding	35
10.2.2	Type 1 AU (Class P)	36
10.2.3	Type 2 AU (Class N)	37
10.2.4	Type 3 AU (Class M)	37
10.2.5	Type 4 AU (Class I)	38
10.2.6	Type 5 AU (Class HM)	40
10.2.7	Type 6 AU (Class U)	40
10.3	dataset_type = 2	40
10.3.1	Type 1 AU	41
10.3.2	Type 2 AU	42
10.3.3	Type 3 AU	42
10.3.4	Type 4 AU	42
10.3.5	Type 6 AU	42
10.4	Genomic descriptors	43
10.4.1	pos	43
10.4.2	rcomp	44
10.4.3	flags	44
10.4.4	mmpos	45
10.4.5	mmtype	47
10.4.6	clips	50
10.4.7	ureads	53
10.4.8	rlen	53
10.4.9	pair	55
10.4.10	mscore	62
10.4.11	mmap	63
10.4.12	msar	66
10.4.13	rtype	66
10.4.14	rgroup	68
10.4.15	qv	68
10.4.16	rname	72
10.4.17	rftp	72
10.4.18	rftt	73
10.4.19	tokentype descriptors	73
10.5	sequence	81
10.5.1	Aligned reads (Classes P, N, M, I, HM)	82
10.5.2	Unmapped reads (Class HM, U)	83
10.6	e-cigar	83
10.6.1	Syntax	83
10.6.2	Decoding process for the first alignment	84
10.6.3	Decoding process for other alignments	92
10.6.4	Reference transformation	92
11	Representation of reference sequences	93
11.1	External reference	94
11.2	Embedded reference	94
11.3	Computed reference	94
11.3.1	General	94
11.3.2	Reference transformation	94
11.3.3	PushIn	95
11.3.4	Local assembly	96
11.3.5	Global assembly	97
12	Block payload parsing process	97
12.1	General	97
12.2	Inverse binarizations	98
12.2.1	Binary (BI)	99

12.2.2	Truncated Unary (TU)	99
12.2.3	Exponential Golomb (EG)	99
12.2.4	Truncated Exponential Golomb (TEG)	100
12.2.5	Signed Truncated Exponential Golomb (STEG)	100
12.2.6	Split Unit-wise Truncated Unary (SUTU)	101
12.2.7	Signed Split Unit-wise Truncated Unary (SSUTU)	101
12.2.8	Double Truncated Unary (DTU)	101
12.2.9	Signed Double Truncated Unary (SDTU)	102
12.3	Decoder configuration	102
12.3.1	Sequences and quality values	102
12.3.2	Support values	103
12.3.3	CABAC binarizations	104
12.3.4	Transformation parameters	107
12.3.5	Msar descriptor and read identifiers	108
12.3.6	State variables	109
12.4	Initialization process for context variables	112
12.5	Arithmetic decoding engine	112
12.5.1	Initialization	112
12.5.2	Arithmetic decoding process	113
12.6	Decoding process for sequence descriptors	120
12.6.1	General	120
12.6.2	Block payload decoding process	121
13	Output format	135
13.1	General	135
13.2	MPEG-G record	135
13.2.1	number_of_template_segments	137
13.2.2	number_of_record_segments	137
13.2.3	number_of_alignments	137
13.2.4	class_ID	137
13.2.5	read_group_len	138
13.2.6	read_1_first	138
13.2.7	seq_ID	138
13.2.8	as_depth	138
13.2.9	read_len	138
13.2.10	qv_depth	138
13.2.11	read_name_len	138
13.2.12	read_name	138
13.2.13	read_group	138
13.2.14	sequence	139
13.2.15	quality_values	139
13.2.16	mapping_pos	139
13.2.17	ecigar_len	139
13.2.18	ecigar_string	139
13.2.19	reverse_comp	139
13.2.20	mapping_score	139
13.2.21	split_alignment	139
13.2.22	delta	140
13.2.23	split_pos	140
13.2.24	split_seq_ID	140
13.2.25	flags	140
13.2.26	more_alignments	140
13.2.27	next_pos	140
13.2.28	next_seq_ID	140
13.3	Initialization process	140
Annex A (informative)	Tokenization of reads identifiers	143
Annex B (informative)	Mapping quality	145
Annex C (informative)	Inverse binarization examples	146