

ISO/IEC 23092-2:2024-03 (E)

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

Contents		Page
Foreword		vii
Introduction		viii
1	Scope	1
2	Normative references	1
3	Terms and definitions	1
4	Abbreviated terms	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	8
5.10	Variables, syntax elements and tables	9
5.11	Text description of logical operators	10
5.12	Processes	11
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	14
7.1	General	14
7.2	Data unit	15
7.3	Raw reference	16
7.3.1	General	16
7.3.2	Syntax and semantics	16
7.4	Parameter set	16
7.4.1	Syntax and semantics	16
7.4.2	Encoding parameters	17
7.5	Access unit	23
7.5.1	Syntax and semantics	24
7.5.2	Access unit types	27
8	Descriptors	28
9	Sequencing reads	31
9.1	General	31
9.2	Supported symbols	31
9.3	Paired-end reads	33
9.4	Reverse-complement reads	33
9.5	Data classes	33
9.6	Aligned data	34
9.7	Unaligned data	35

10	Decoding process	35
10.1	General.....	35
10.2	dataset_type = 0 or 1.....	36
10.2.1	General.....	36
10.2.2	References padding.....	36
10.2.3	Type 1 AU (Class P).....	37
10.2.4	Type 2 AU (Class N).....	38
10.2.5	Type 3 AU (Class M).....	38
10.2.6	Type 4 AU (Class I).....	39
10.2.7	Type 5 AU (Class HM).....	41
10.2.8	Type 6 AU (Class U).....	41
10.3	dataset_type = 2.....	42
10.3.1	General.....	42
10.3.2	Type 1 AU.....	43
10.3.3	Type 2 AU.....	43
10.3.4	Type 3 AU.....	43
10.3.5	Type 4 AU.....	44
10.3.6	Type 6 AU.....	44
10.4	Genomic descriptors.....	44
10.4.1	General.....	44
10.4.2	pos.....	45
10.4.3	rcomp.....	45
10.4.4	flags.....	46
10.4.5	mmpos.....	47
10.4.6	mmtree.....	49
10.4.7	clips.....	53
10.4.8	ureads.....	55
10.4.9	rlen.....	56
10.4.10	pair.....	57
10.4.11	mscore.....	64
10.4.12	mmap.....	65
10.4.13	msar.....	68
10.4.14	rtype.....	69
10.4.15	rgroup.....	71
10.4.16	qv.....	71
10.4.17	rname.....	75
10.4.18	rftp.....	75
10.4.19	rftt.....	76
10.4.20	tokentype descriptors.....	77
10.5	sequence.....	85
10.5.1	General.....	85
10.5.2	Aligned reads (Classes P, N, M, I, HM).....	85
10.5.3	Unmapped reads (Class HM, U).....	86
10.6	e-cigar.....	86
10.6.1	Syntax.....	86
10.6.2	Decoding process for the first alignment.....	88
10.6.3	Decoding process for other alignments.....	95
10.6.4	Reference transformation.....	95
11	Representation of reference sequences	96
11.1	External reference.....	97
11.2	Embedded reference.....	97
11.3	Computed reference.....	97
11.3.1	General.....	97
11.3.2	Supported Algorithms.....	97
11.3.3	Reference transformation.....	98
11.3.4	PushIn.....	98
11.3.5	Local assembly.....	100
11.3.6	Global assembly.....	101
12	Block payload parsing process	101
12.1	General.....	101
12.2	Encoding Mode 0.....	102
12.3	Inverse binarizations.....	103
12.3.1	General.....	103
12.3.2	Binary (BI).....	103

12.3.3	Truncated unary (TU)	104
12.3.4	Exponential golomb (EG)	104
12.3.5	Truncated exponential golomb (TEG)	105
12.3.6	Signed truncated exponential golomb (STEG)	105
12.3.7	Split unit-wise truncated unary (SUTU)	105
12.3.8	Signed split unit-wise truncated unary (SSUTU)	106
12.3.9	Double truncated unary (DTU)	106
12.3.10	Signed double truncated unary (SDTU)	107
12.4	Decoder configuration	107
12.4.1	Sequences and quality values	107
12.4.2	Support values	108
12.4.3	CABAC binarizations	109
12.4.4	Transformation parameters	112
12.4.5	Msar descriptor and read identifiers	113
12.4.6	State variables	114
12.5	Initialization process for context variables	117
12.6	Arithmetic decoding engine	117
12.6.1	Initialization	117
12.6.2	Arithmetic decoding process	118
12.7	Decoding process for sequence descriptors	124
12.7.1	General	125
12.7.2	Block payload decoding process	125
12.8	BSC decoding process	139
12.8.1	decoding process	139
Output format		141
13.1	General	141
13.2	MPEG-G record	141
13.2.1	General	141
13.2.2	number_of_template_segments	143
13.2.3	number_of_record_segments	143
13.2.4	number_of_alignments	143
13.2.5	class_ID	144
13.2.6	read_group_len	144
13.2.7	reserved	144
13.2.8	read_1_first	144
13.2.9	seq_ID	144
13.2.10	as_depth	144
13.2.11	read_len	144
13.2.12	qv_depth	144
13.2.13	read_name_len	144
13.2.14	read_name	145
13.2.15	read_group	145
13.2.16	sequence	145
13.2.17	quality_values	145
13.2.18	mapping_pos	145
13.2.19	ecigar_len	145
13.2.20	ecigar_string	145
13.2.21	reverse_comp	145
13.2.22	mapping_score	145
13.2.23	split_alignment	145
13.2.24	delta	146
13.2.25	split_pos	146
13.2.26	split_seq_ID	146
13.2.27	flags	146
13.2.28	more_alignments	146
13.2.29	next_pos	146
13.2.30	next_seq_ID	146
13.3	Initialization process	146
Annex A (informative) Tokenization of reads identifiers		150
Annex B (informative) Mapping quality		152
Annex C (informative) Inverse binarization examples		153
Annex D Block Sorting, Lossless Data Compression		157