**SAM FLAG field**

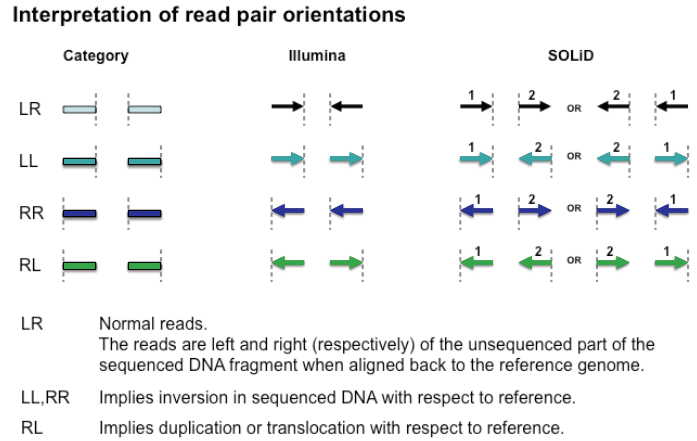
Strictly speaking, 0x10 (16 in decimal) in the FLAG field means “SEQ being reverse complemented” [Based on SAM spec v 1.4]. However, Picard’s “SAM FLAG explainer” describes 0x10 as meaning, “read reverse strand”. Similarly, 0x20 strictly means “SEQ of next segment in the template being reversed”, but Picard describes this as “mate reverse strand”.

**Illumina paired-end reads**

Read1 and read2 are from opposite strands of the same DNA molecule (<http://seqanswers.com/forums/showthread.php?t=18158>).

**Read orientation in Novoalign DNA-seq SAM files**

I believe this also holds for Bowtie and BWA SAM files.



(From <http://www.broadinstitute.org/software/igv/interpreting_pair_orientations>)

**IGV**

<table><tr><td valign="top"><b>Left alignment</b><br/>Sample = ASD2-41

Read group = 1

----------------------

Read name = HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT

Alignment start = 19558864 (+)

Cigar = 109M1H

Mapped = yes

Mapping quality = 12

----------------------

----------------------

Pair start = chr14:19558989 (-)

Pair is mapped = yes

Insert size = 234

Pair orientation = F2R1

----------------------

Second in pair

-------------------

LB = unknown

MD = 109

PG = novoalign

RG = 1

AM = 150

NM = 0

SM = 150

PQ = 1

UQ = 0

AS = 0

PU = HWI-ST0798\_0062

-------------------</td><td valign="top"><b>Right alignment</b><br/>Sample = ASD2-41

Read group = 1

----------------------

Read name = HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT

Alignment start = 19558989 (-)

Cigar = 110M

Mapped = yes

Mapping quality = 12

----------------------

Base = G

Base phred quality = 38

----------------------

Pair start = chr14:19558864 (+)

Pair is mapped = yes

Insert size = -234

Pair orientation = F2R1

----------------------

First in pair

-------------------

LB = unknown

MD = 110

PG = novoalign

RG = 1

AM = 150

NM = 0

SM = 150

PQ = 1

UQ = 1

AS = 1

PU = HWI-ST0798\_0062

-------------------</td></tr></table>

Alignment start position = chr14:19558864

Null

**BAM**

HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT 163 chr14 19558864 12 109M1H = 19558989 234 GGGTTTGATTTCAATACATAGCATAAAAATGAGTTTTCTCCTTTAAATATAACTAGTTGGTGAAAGCTGTGGAATGTTATTTTGAAATCCTAGGATTTGTAATTTGTTT AAAAABDBCCCFCCDCGCDCFGCDCDDDDDGDGDDDDGDGGDDDDDDEDEDDGEDFEDFFDFDDDGGEG@=FADEGEECEEEEG?ABDFFEB>AADDDBBCCDCCBCB> LB:Z:unknown MD:Z:109 PG:Z:novoalign RG:Z:1 AM:i:150 NM:i:0 SM:i:150 PQ:i:1 UQ:i:0 AS:i:0 PU:Z:HWI-ST0798\_0062

HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT 83 chr14 19558989 12 110M = 19558864 -234 AGGCCGTACAATGCCGGGAAGATGAATGTGCGTTAATGTTGCTGGAACATGGCACTGATCCGAATATTCCAGATGAGTATGGAAATACCGCTCTACACTATGCTATCTAC @A@>@ECCCCCAEADEEFCCFCCEBBDFDGFGEEDDEGEEGGEGGDDGDEGGGDGEGDEGGGDDEDEEGGDGDEGDGECEGGCCCB8FFGGDFDCGCFDCDEFCBBCA@@ LB:Z:unknown MD:Z:110 PG:Z:novoalign RG:Z:1 AM:i:150 NM:i:0 SM:i:150 PQ:i:1 UQ:i:1 AS:i:1 PU:Z:HWI-ST0798\_0062

**FASTQ**

@HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT/1

GTAGATAGCATAGTGTAGAGCGGTATTTCCATACTCATCTGGAATATTCGGATCAGTGCCATGTTCCAGCAACATTAACGCACATTCATCTTCCCGGCATTGTACGGCCT

+HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT/1

aaaeeeeeggggfhhgfhhiifiRbfhiiiiiiiiiiiiiiihiiiiiiiiiiiiihiihihifffghhhiigggdgeeeecccbbb`ccccdcccac\_cccdcda\_\_aa

@HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT/2

GGGTTTGATTTCAATACATAGCATAAAAATGAGTTTTCTCCTTTAAATATAACTAGTTGGTGAAAGCTGTGGAATGTTATTTTGAAATCCTAGGATTTGTAATTTGTTTT

+HWI-ST0798\_0086:3:2105:9728:41512#TAGCTT/2

bbbeeeeegggggiiiiiiihiiiiihhhiihiehffhiiiihihiiiiiihhhhhghiieghhhhhhh\_]e`gfhiidhgfgf\`adddeb\_b\_ddcb`cdecccdc^B

**Read orientation in Bismark BS-seq SAM files**

**IGV**

<b>Left alignment</b><br/>----------------------

Read name = SRR097428.6814839\_HWI-BRUNOP20X:0637:1:5:9089:67292\_length=75/1

Alignment start = 19011703 (+)

Cigar = 75M

Mapped = yes

Mapping quality = 255

----------------------

Base = A

Base phred quality = 39

----------------------

Pair start = chr14:19011768 (+)

Pair is mapped = yes

Insert size = 140

Pair orientation = F1F2

----------------------

First in pair

-------------------

XG = CT

NM = 14

XM = .x..h.h................h..h.x.............h...h....h..........x.......

.hh.x

XR = CT

XX = 1C2C1C16C2C1C13C3C3AC10C8CC1C

-------------------

Alignment start position = chr14:19011703

null

<b>Left alignment</b><br/>----------------------

Read name = SRR097428.6814839\_HWI-BRUNOP20X:0637:1:5:9089:67292\_length=75/2

Alignment start = 19011768 (+)

Cigar = 75M

Mapped = yes

Mapping quality = 255

----------------------

Base = A

Base phred quality = 31

----------------------

Pair start = chr14:19011703 (+)

Pair is mapped = yes

Insert size = -140

Pair orientation = F1F2

----------------------

Second in pair

-------------------

XG = CT

NM = 16

XM = ......hh.x...h......................x.....hh..Z......h.h....h...hh...h

..hx.

XR = GA

XX = 6CC1C3C22C2T2CC9C1C4C3CC3C2CC1

-------------------

Alignment start position = chr14:19011768

Null

**BAM**

SRR097428.6814839\_HWI-BRUNOP20X:0637:1:5:9089:67292\_length=75/1 67 chr14 19011703 255 75M = 19011768 140

ATTGTTTTATGAAAAGGAATGTTTAATTTTGTGAGTTGAATGTAAGTATGGTAAAAAAGTTTTTGAGAATGTTTT HHHHHHHGGHHHHHHHHIHHHHHHHHHHHHHHHGHEHHFHHHIHHHDHHHHEEEHF9FF<D@CDFGHGGDHAEHF NM:i:14 XX:Z:1C2C1C16C2C1C13C3C3AC10C8CC1C XM:Z:.x..h.h................h..h.x.............h...h....h..........x........hh.x XR:Z:CT XG:Z:CT

SRR097428.6814839\_HWI-BRUNOP20X:0637:1:5:9089:67292\_length=75/2 131 chr14 19011768 255 75M = 19011703 -140 AGAATGTTTTTGTTTAGTTTTTATTTGAATATAATATTGNTTTTAACGAAAGGTTTAAAGTTTTTTAAATATTTA DGF@908/;(EGFDGDGGB@FDFFEHHFHGDADADDADD!A@A@?HBHHHHHHEHHHHHGHHHHHHHHHEHHHHH NM:i:16 XX:Z:6CC1C3C22C2T2CC9C1C4C3CC3C2CC1 XM:Z:......hh.x...h......................x.....hh..Z......h.h....h...hh...h..hx. XR:Z:GA XG:Z:CT

**FASTQ**

@SRR097428.6814839 HWI-BRUNOP20X:0637:1:5:9089:67292 length=75

ATTGTTTTATGAAAAGGAATGTTTAATTTTGTGAGTTGAATGTAAGTATGGTAAAAAAGTTTTTGAGAATGTTTT

+SRR097428.6814839 HWI-BRUNOP20X:0637:1:5:9089:67292 length=75

HHHHHHHGGHHHHHHHHIHHHHHHHHHHHHHHHGHEHHFHHHIHHHDHHHHEEEHF9FF<D@CDFGHGGDHAEHF

@SRR097428.6814839 HWI-BRUNOP20X:0637:1:5:9089:67292 length=75

TAAATATTTAAAAAACTTTAAACCTTTCGTTAAAANCAATATTATATTCAAATAAAAACTAAACAAAAACATTCT

+SRR097428.6814839 HWI-BRUNOP20X:0637:1:5:9089:67292 length=75

HHHHHEHHHHHHHHHGHHHHHEHHHHHHBH?@A@A!DDADDADADGHFHHEFFDF@BGGDGDFGE(;/809@FGD

**Bismark translation**

If we know that the CTOB and CTOT are merely theoretical, then we can use the “correct” strand information in the FLAG for each read and still uniquely determine whether a paired-read is informative for the OT or OB by checking the FLAG of the first read in the pair. If the strand of read1 is “+“ then the paired-read is informative for the OT; if the strand of read1 is “-“ then the paired-read is informative for the OB.

My proposal to resolve conflicts around the use of the FLAG field between the official SAM spec and Bismark’s usage is as follows:

* Encode the “correct” strand information in the FLAG, i.e. +/- (0x20, 0x10) or -/+ (0x10, 0x20).
* To resolve ambiguities as to whether a read pair with +/- is informative for the OT or CTOB and whether a read pair with -/+ is informative for the OB or CTOT, encode this as a TAG XS:Z:OT, XS:Z:CTOB, XS:Z:CTOT, XS:Z:OB.

In this way the Bismark SAM files will conform to the SAM specifications (v1.4) while incorporating information about which of the four possible bisulfite strands of DNA the read originated from.

Table 1: Translation of Bismark strand flags for paired-end reads (read1/read2) for each of the four possible bisulfite strands of DNA. OT=original top; CTOB=complementary to original bottom; CTOT=complementary to original top; OB=original bottom. Only reads from the OT and OB strands are theoretically possible when using Illumina’s non-directional BS-seq library. “Correct” strand refers to the orientation of normal paired-end Illumina reads – IGV expects a paired to have this orientation in order to treat the read as a “correct” pair. Table partially based on information posted by Felix Krueger on seqanswers.com (<http://seqanswers.com/forums/showthread.php?t=18422>)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Read informative for** | **Read conversion** | **Genome conversion** | **“Correct” strand** | **Bismark strand** | **Correct FLAG values** | **XS tag** |
| **OT** | CT | CT | +/- | +/+ | 0x20/0x10 | XS:Z:OT |
| **CTOB** | GA | GA | +/- | -/- | 0x20/0x10 | XS:Z:CTOB |
| **CTOT** | GA | CT | -/+ | +/+ | 0x10/0x20 | XS:Z:CTOT |
| **OB** | CT | GA | -/+ | -/- | 0x10/0x20 | XS:Z:OB |

This information can be inferred from the XR and XG tags (encoding “read conversion” and “genome conversion” information) of Bismark’s current SAM files.

I will write a pysam script to correct the strand information and add the XS tag to my files.