Next Generation Sequencing File Formats.

Pierre Lindenbaum
@yokofakun
pierre.lindenbaum@univ-nantes.fr
http://plindenbaum.blogspot.com
https://github.com/lindenb/courses

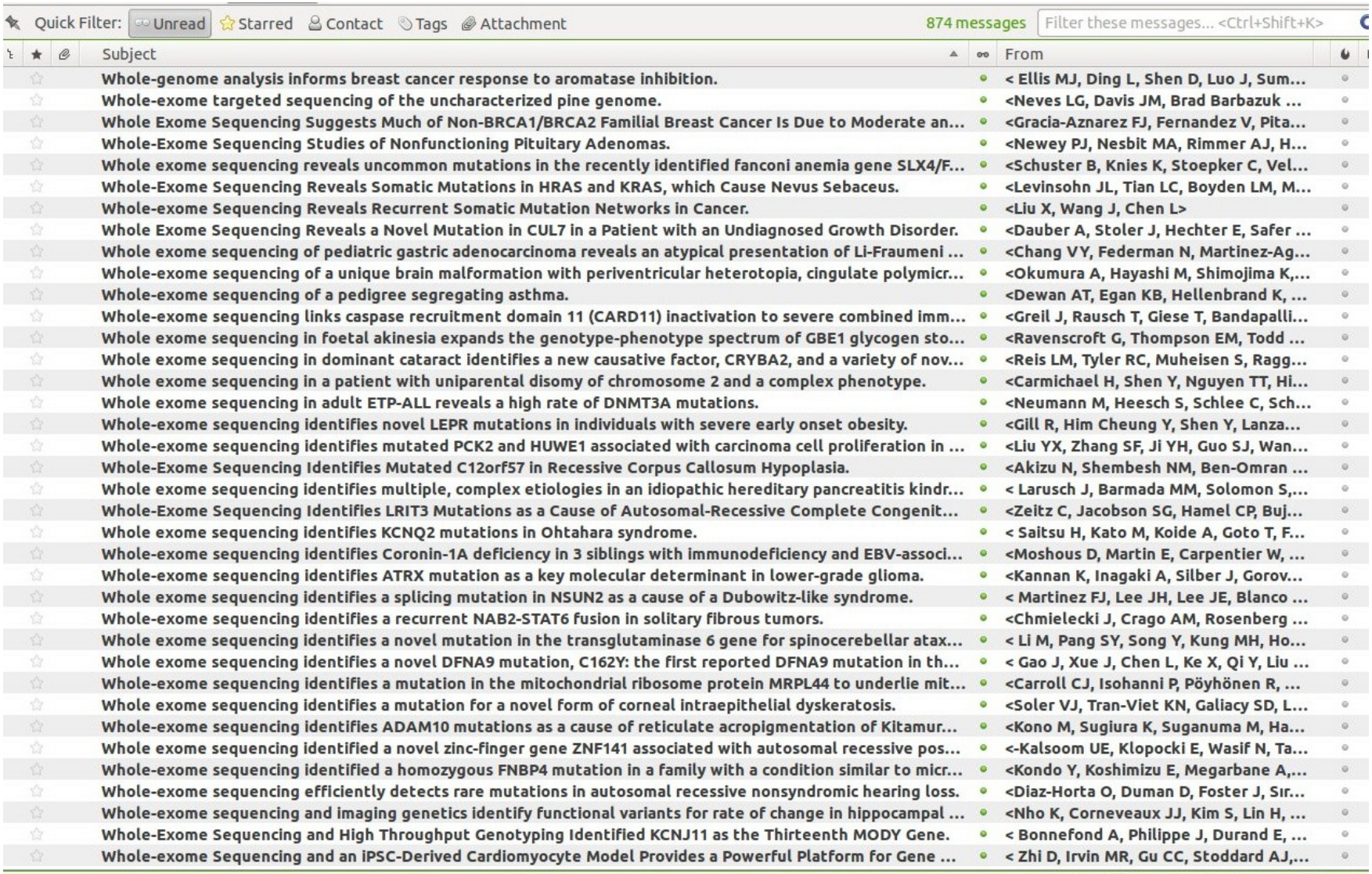
Institut du Thorax. Nantes. France

September 23, 2013

You don't need to have a deep knowledge of those formats.
(Unless you're doing NGS)

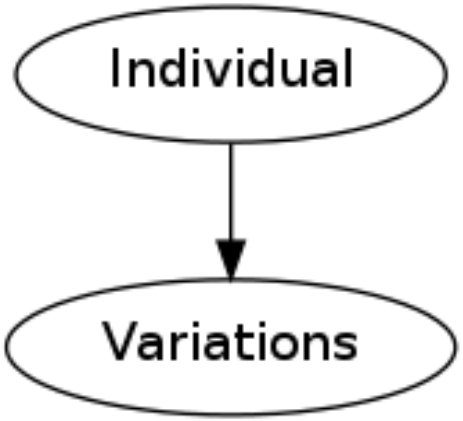
Understand how people have solved their BIG data problems.

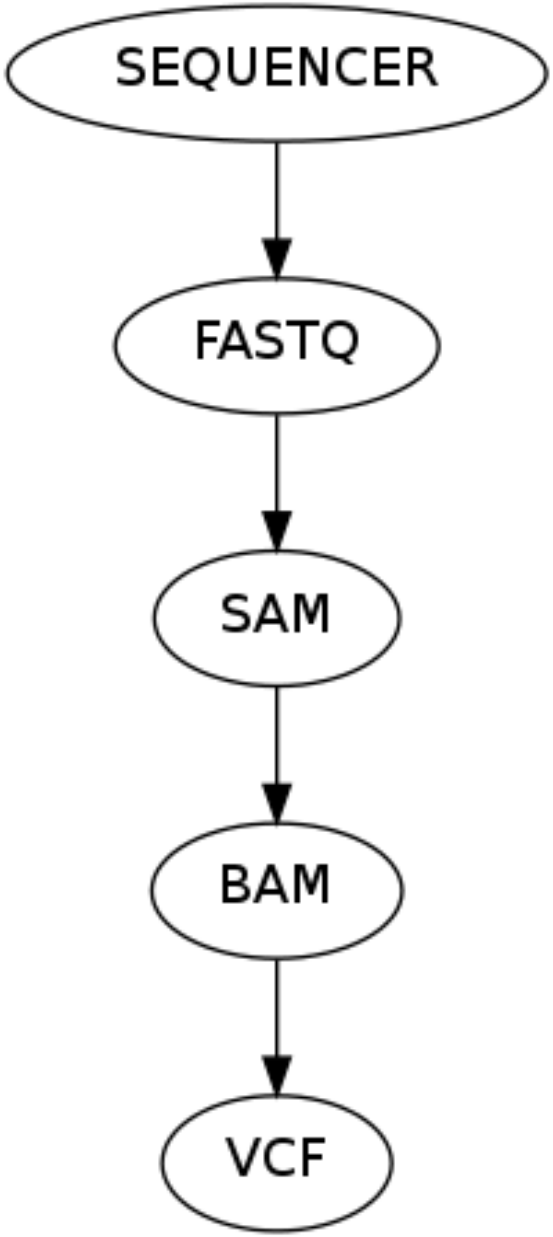
Why sequencing ?



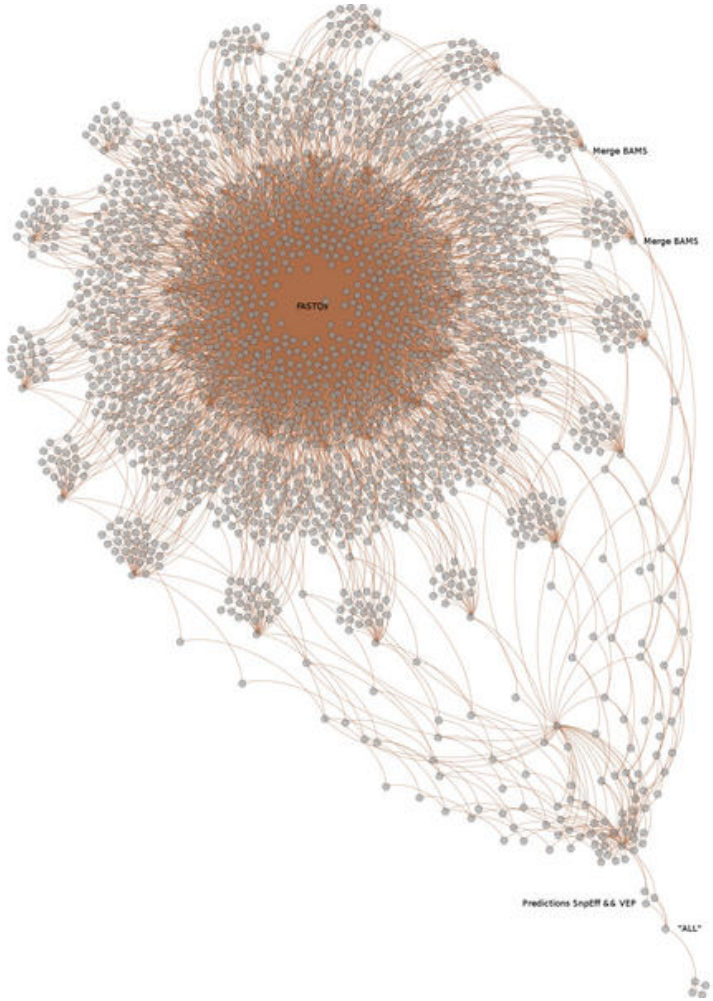
Quick Filter: Unread Starred Contact Tags Attachment — 874 messages — Filter these messages... <Ctrl+Shift+K>

Subject	From
Whole-genome analysis informs breast cancer response to aromatase inhibition.	< Ellis MJ, Ding L, Shen D, Luo J, Sum...
Whole-exome targeted sequencing of the uncharacterized pine genome.	<Neves LG, Davis JM, Brad Barbazuk ...
Whole Exome Sequencing Suggests Much of Non-BRCA1/BRCA2 Familial Breast Cancer Is Due to Moderate an...	<Gracia-Aznarez FJ, Fernandez V, Pita...
Whole-Exome Sequencing Studies of Nonfunctioning Pituitary Adenomas.	<Newey PJ, Nesbit MA, Rimmer AJ, H...
Whole exome sequencing reveals uncommon mutations in the recently identified fanconi anemia gene SLX4/F...	<Schuster B, Knies K, Stoepker C, Vel...
Whole-Exome Sequencing Reveals Somatic Mutations in HRAS and KRAS, which Cause Nevus Sebaceus.	<Levinsohn JL, Tian LC, Boyden LM, M...
Whole-exome Sequencing Reveals Recurrent Somatic Mutation Networks in Cancer.	<Liu X, Wang J, Chen L>
Whole Exome Sequencing Reveals a Novel Mutation in CUL7 in a Patient with an Undiagnosed Growth Disorder.	<Dauber A, Stoler J, Hechter E, Safer ...
Whole exome sequencing of pediatric gastric adenocarcinoma reveals an atypical presentation of Li-Fraumeni ...	<Chang VY, Federman N, Martinez-Ag...
Whole-exome sequencing of a unique brain malformation with periventricular heterotopia, cingulate polymicr...	<Okumura A, Hayashi M, Shimojima K,...
Whole-exome sequencing of a pedigree segregating asthma.	<Dewan AT, Egan KB, Hellenbrand K, ...
Whole-exome sequencing links caspase recruitment domain 11 (CARD11) inactivation to severe combined imm...	<Greil J, Rausch T, Giese T, Bandapalli...
Whole exome sequencing in foetal akinesia expands the genotype-phenotype spectrum of GBE1 glycogen sto...	<Ravenscroft G, Thompson EM, Todd ...
Whole exome sequencing in dominant cataract identifies a new causative factor, CRYBA2, and a variety of nov...	<Reis LM, Tyler RC, Muheisen S, Ragg...
Whole exome sequencing in a patient with uniparental disomy of chromosome 2 and a complex phenotype.	<Carmichael H, Shen Y, Nguyen TT, Hi...
Whole exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations.	<Neumann M, Heesch S, Schlee C, Sch...
Whole-exome sequencing identifies novel LEPR mutations in individuals with severe early onset obesity.	<Gill R, Him Cheung Y, Shen Y, Lanza...
Whole-exome sequencing identifies mutated PCK2 and HUWE1 associated with carcinoma cell proliferation in ...	<Liu YX, Zhang SF, Ji YH, Guo SJ, Wan...
Whole-Exome Sequencing Identifies Mutated C12orf57 in Recessive Corpus Callosum Hypoplasia.	<Akizu N, Shembesh NM, Ben-Omran ...
Whole exome sequencing identifies multiple, complex etiologies in an idiopathic hereditary pancreatitis kindr...	< Larusch J, Barmada MM, Solomon S,...
Whole-Exome Sequencing Identifies LRIT3 Mutations as a Cause of Autosomal-Recessive Complete Congenit...	<Zeitz C, Jacobson SG, Hamel CP, Buj...
Whole exome sequencing identifies KCNQ2 mutations in Ohtahara syndrome.	< Saitsu H, Kato M, Koide A, Goto T, F...
Whole-exome sequencing identifies Coronin-1A deficiency in 3 siblings with immunodeficiency and EBV-associ...	<Moshous D, Martin E, Carpentier W, ...
Whole-exome sequencing identifies ATRX mutation as a key molecular determinant in lower-grade glioma.	<Kannan K, Inagaki A, Silber J, Gorov...
Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome.	< Martinez FJ, Lee JH, Lee JE, Blanco ...
Whole-exome sequencing identifies a recurrent NAB2-STAT6 fusion in solitary fibrous tumors.	<Chmielecki J, Crago AM, Rosenberg ...
Whole exome sequencing identifies a novel mutation in the transglutaminase 6 gene for spinocerebellar atax...	< Li M, Pang SY, Song Y, Kung MH, Ho...
Whole exome sequencing identifies a novel DFNA9 mutation, C162Y: the first reported DFNA9 mutation in th...	< Gao J, Xue J, Chen L, Ke X, Qi Y, Liu ...
Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mit...	<Carroll CJ, Isohanni P, Pöyhönen R, ...
Whole exome sequencing identifies a mutation for a novel form of corneal intraepithelial dyskeratosis.	<Soler VJ, Tran-Viet KN, Galiacy SD, L...
Whole-exome sequencing identifies ADAM10 mutations as a cause of reticulate acropigmentation of Kitamur...	<Kono M, Sugiura K, Suganuma M, Ha...
Whole exome sequencing identified a novel zinc-finger gene ZNF141 associated with autosomal recessive pos...	<-Kalsoom UE, Klopocki E, Wasif N, Ta...
Whole-exome sequencing identified a homozygous FNBP4 mutation in a family with a condition similar to micr...	<Kondo Y, Koshimizu E, Megarbane A,...
Whole-exome sequencing efficiently detects rare mutations in autosomal recessive nonsyndromic hearing loss.	<Diaz-Horta O, Duman D, Foster J, Sır...
Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal ...	<Nho K, Corneveaux JJ, Kim S, Lin H, ...
Whole-Exome Sequencing and High Throughput Genotyping Identified KCNJ11 as the Thirteenth MODY Gene.	< Bonnefond A, Philippe J, Durand E, ...
Whole-exome Sequencing and an iPSC-Derived Cardiomyocyte Model Provides a Powerful Platform for Gene ...	< Zhi D, Irvin MR, Gu CC, Stoddard AJ,...


Individual
Variations


SEQUENCER
FASTQ
SAM
BAM
VCF

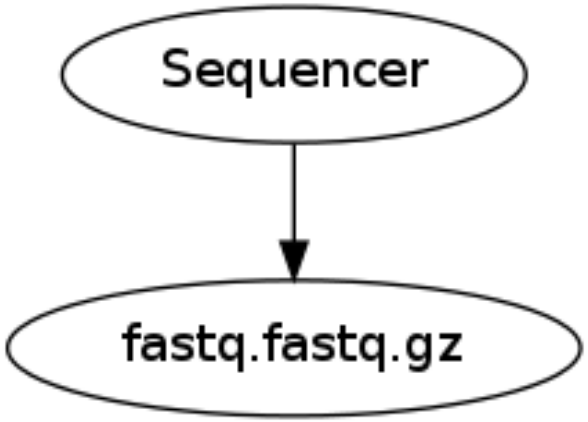
Well, that's a little more complicated ...



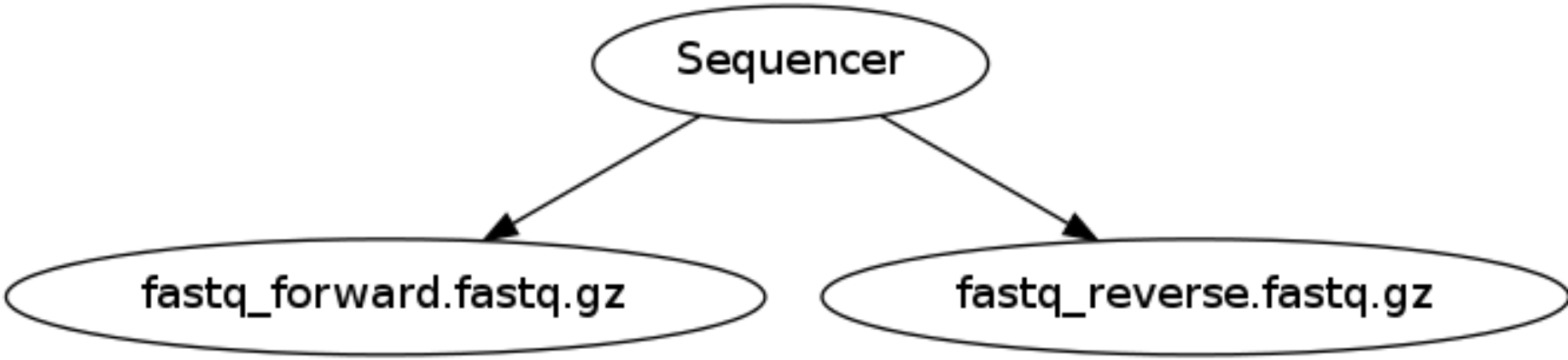
FASTQ

FASTQ

FASTQ: text-based format for storing both a DNA sequence and its corresponding quality scores


Sequencer
fastq.fastq.gz

FASTQ
FASTQ for paired end



FASTQ Example

```
@IL31_4368:1:1:996:8507/1
NTGATAAAGTAATGACAAAATAATGACATTATTGTTACTATGGTTACTGTGGGA
+
(94**0-)*7=06>>><<<<<<22@>6;;;5;6:;63:4?-622647..-.5.%
@IL31_4368:1:1:996:21421/1
NAAGTTAATTCTTCATTGTCCATTCCTCTGAAATGATTCAGAAATACTGGTAGT
+
(***+*2396,@<+<:@@@;;5)<0)69606>4;5>;>6&<102)0*+8:&137;
@IL31_4368:1:1:997:10572/1
NAATGTATGTAGACCCTTCACATTCAAAGGCAAATACAATATCATCATGTCTTC
+
(/9**-0032>:>>9>4@@=>??@@:-66,;>;<;6+;255,1;7>>>>3676'
@IL31_4368:1:1:997:15684/1
NGCAATCAATGCTATGATTGATCCTGATGGAACTTTGGAGGCTCTGAACAACAT
+
()1,*37766>@@@>?@<?@@:>@0>>><-888>8;>*;966>;;;@8@4,.2.
@IL31_4368:1:1:997:15249/1
```

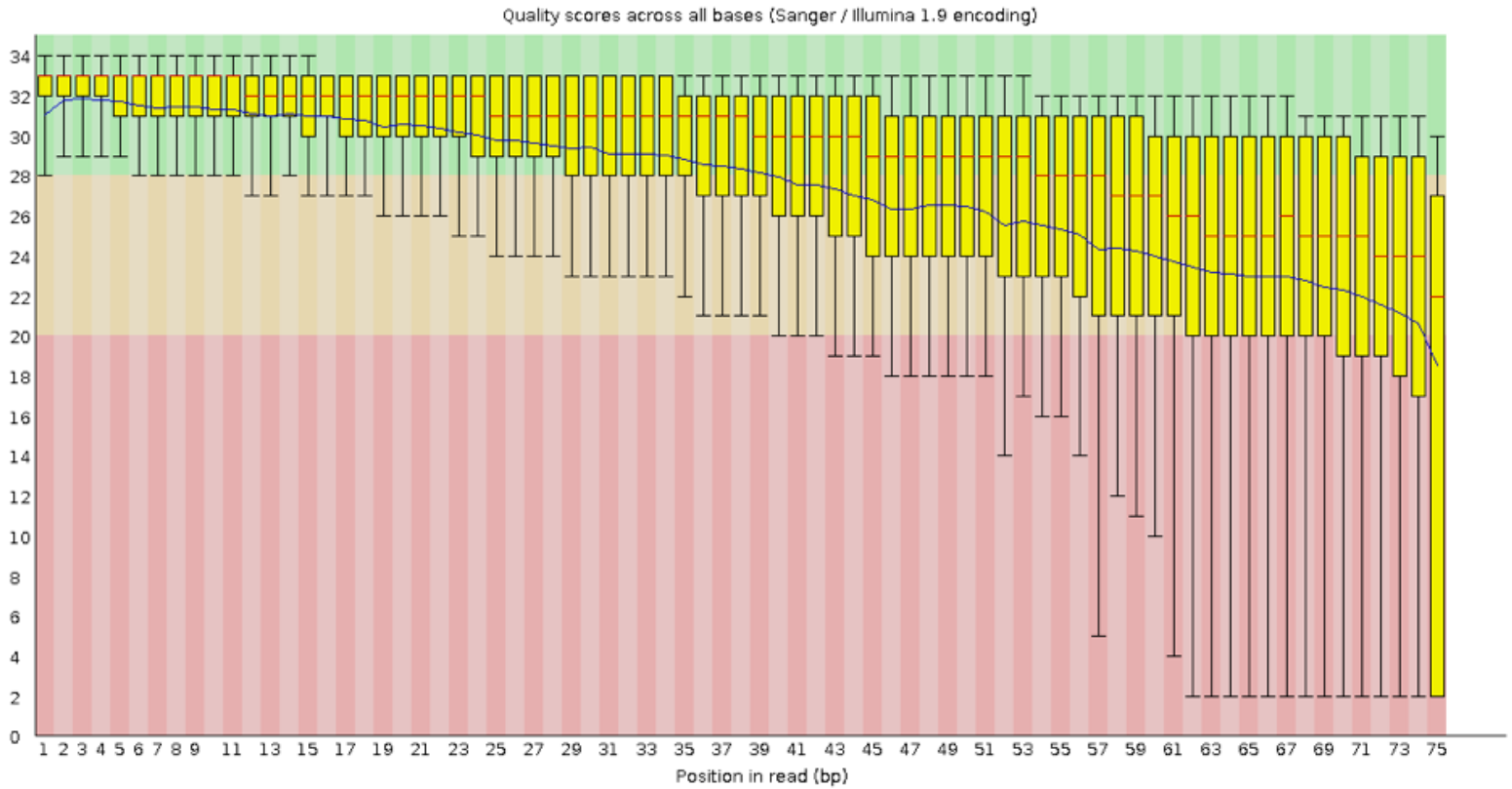
FASTQ name

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

Col	Brief description
EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even num
ATCACG	index sequence

FASTQ Quality

Per base sequence quality



FASTQ Quality

A quality value Q is an integer mapping of p (i.e., the probability that the corresponding base call is incorrect).

$Q_{\text{sanger}} = -10 \log_{10} p$

FASTQ Quality

Since a human readable format is desired for SAM, 33 is added to the calculated quality in order to make it a printable character ranging from ! - .

$$Q_{\text{sanger}} = -10 \log_{10} p + 33$$

Aligned Reads

```
44187101  44187111  44187121  44187131  44187141  44187151  44187161  44187171
aaatgagccaggtgtggtggtgcacacctatagtcccagctacgcaggaggctgaggtgggaggatcgcttaaacccggc REFERENCE
.........................................Y..................................... CONSENSUS
aaa gagccaggtgtggtggtgcacaccgataggcccagctacgtaggaggctgaggtgggaggatcgcttaaa  cggc
AAA GAGCCAGGTGTGGTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAA  CGGC
aaatga CCAGGTGTGGTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAACCC  c
aaatgagcc GGTGTGGTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAACCCGGC
AAATGAGCCAGG gtggtggtgcacacctatagtcccagcgacgtaggaggctgaggtgggaggatcgcttaaacccggc
AAATGAGCCAGGTG ggtggtgcacacctatagtcccagctaagtaggaggctgaggtgggaggatcgctttaacccggc
AAATGAGCCAGGTGT GTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAACCGGGC
ACATGAGCCAGGTGTG tggtgcacacctatagtcccagctacgtaggaggctgaggtgggaggatcgcttaaacccggc
aaatgagccaggtgtgg    GCACACGTAAAGTCCCAGCTACGCAGGAGGCTGAGGTGGGAGGATCGCTTAAACCCGGC
CAATGAGCCAGTTGTGG     cacacctatagtcccagctacgcacgaggctgaggtgggaggatcgctttaacccggc
AAATGAGCCAGGTGAGGT    cacacctatagtcccagctacgcaggaggctgaggtgggaggatcgcttaaacccggc
AAATGAGCCAGGTGTGGT     acacctatagtcccagctacgcaggaggctgaggtgggaggatcgctttaacccggc
aaatgagccaggtgtggtgg      cctatagtcccagctacgtaggaggctgaggtgggaggatcgcttaaacccggc
AAATGAGCCAGGTGTGGTGG        TATAGTCCCAGCTACGCAGGAGGCTGAGGTGGTAGGATCGCATAAACCCGGC
AAATGAGCCAGGTGTGGTGGT         TAGTCCCAGCTACGTAGGAGGCTGAGTTGGGAGGATCTCTTAAACCCGGC
aaatgagccaggtgtggtggtg        TCGTCCCAGCTACGCAGGAGGCTTAGGTGGGAGGATCGCTTAAACCCGGC
aaatgagccaggtgtggtggtgca       AGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGGTTAAACCCGGC
aaatgagccaggtgtggtggtgcac         cccagctacgcaggaggctgaggtgggaccatcgcttaaaccccgc
aaatgagccaggtgtggtggtgcac          CCAGCTACGTAGTAGGCTGAGGTGGGAGGATCGCTTAAACCCGGC
```

SAM

SAM Format

"SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments"

SAM Format

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Format
Structure

```
+ HEADER
   -version
   -program parameters
   +GENOME
    - chrom1 size
    - chrom2 size
    - chrom3 size
    - (..)
   +GROUPS
    - group1 : sample1, lane 4
    - group2 : sample2, lane 1
+ BODY
  - READ1 -> group1
  - READ2 -> group1
  - READ3 -> group1
  - READ4 -> group2
```

SAM Header Section

SAM Example

Simple example

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 83 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

SAM Header

```
@HD VN:1.0 SO:coordinate
@SQ SN:1 LN:249250621 AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta M5:1b22b98cdeb4a9304cb
@SQ SN:2 LN:243199373 AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta M5:a0d9851da00400dec1(
@SQ SN:3 LN:198022430 AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta M5:fdfd811849cc2fadebc
@RG ID:UM0098:1 PL:ILLUMINA PU:HWUSI-EAS1707-615LHAAXX-L001 LB:80 DT:2010-05-05T20:00:00-0400 SM:SD3774
@RG ID:UM0098:2 PL:ILLUMINA PU:HWUSI-EAS1707-615LHAAXX-L002 LB:80 DT:2010-05-05T20:00:00-0400 SM:SD3774
@PG ID:bwa VN:0.5.4
@PG ID:GATK TableRecalibration VN:1.0.3471 CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, Cy
```

SAM Alignment Section

SAM Example

Simple example

```
IL31_4368:1:1:996:8507 77 * 0 0 * * 0 0 NTGATAAAGTAATGACAAAATAATGACATTATTGTTACTATGGTTACTGTGGGA (94**0-)
IL31_4368:1:1:996:8507 141 * 0 0 * * 0 0 TCCCTTACCCCCAAGCTCCATACCCTCCTAATGCCCACACCTCTTACCTTAGGA FFCEFFF
IL31_4368:1:1:996:21421 77 * 0 0 * * 0 0 NAAGTTAATTCTTCATTGTCCATTCCTCTGAAATGATTCAGAAATACTGGTAGT (***+23
IL31_4368:1:1:996:21421 141 * 0 0 * * 0 0 CAAAAACTTTCACTTTACCTGCCGGGTTTCCCAGTTTACATTCCACTGTTTGAC >DBDDF
IL31_4368:1:1:997:10572 77 * 0 0 * * 0 0 NAATGTATGTAGACCCTTCACATTCAAAGGCAAATACAATATCATCATGTCTTC (/9**-(
IL31_4368:1:1:997:10572 141 * 0 0 * * 0 0 GATCTTCTGTGACTGGAAGAAAATGTGTTACATATTACATTTCTGTCCCCATTG E?=EEC
IL31_4368:1:1:997:15684 83 chr1 241356612 60 54M = 241356442 -224 ATGTTGTTCAGAGCCTCCAAAGTTCCATCAGGATCAA
IL31_4368:1:1:997:15684 163 chr1 241356442 60 54M = 241356612 224 CAGCCTCAGATTCAGCATTCTCAAATTCAGCTGCGGC
IL31_4368:1:1:997:15249 77 * 0 0 * * 0 0 NCGTTATAATGGAATTATTTTTCTTCCTTTATTTAATGTGTTGACAAAGAGAAC (916928
IL31_4368:1:1:997:15249 141 * 0 0 * * 0 0 AATGTTCTGAAACCTCTGAGAAAGCAAATATTTATTTTAATGAAAAATCCTTAT EDEEC;
IL31_4368:1:1:997:6273 77 * 0 0 * * 0 0 NTACGAAGAAGTATTTCATTGGGAGGAGCTTATCCAAATATTTCCTGTCTATCC (**4*5-*
IL31_4368:1:1:997:6273 141 * 0 0 * * 0 0 ACATTTACCAAGACCAAAGGAAACTTACCTTGCAAGAATTAGACAGTTCATTTG EEAAFFF
IL31_4368:1:1:997:1657 83 chr1 143630364 60 54M = 143630066 -352 TACCTTTTTAAAGAGATCTAAAATTGTCACATGGTTAT
IL31_4368:1:1:997:1657 163 chr1 143630066 60 54M = 143630364 352 CCCACCTCTCTCAATGTTTTCCATATGGCAGGGACTCA
IL31_4368:1:1:997:5609 77 * 0 0 * * 0 0 NGGTGTCTCTTACGGACAGCATTAAGCTAGATTCTTTTTAGACCGATCTGCCAA (*+*&,14
IL31_4368:1:1:997:5609 141 * 0 0 * * 0 0 TCACTATCAGAAACAGAATGTATAACTTCCAAATCAGTAGGAAACACAAGGAAA AEECECF
IL31_4368:1:1:997:14262 77 * 0 0 * * 0 0 NGAGAACCAATGGGAAGCAGCCTGAGCTGCTGGAACCTATTCCCCATGACTTCA (913624
IL31_4368:1:1:997:14262 141 * 0 0 * * 0 0 TGTTTTTTCTTTTTCTTTTTTTTTTGACAGTGCAGAGATTTTTTATCTTTTTAA 97’<2<
IL31_4368:1:1:998:19914 77 * 0 0 * * 0 0 NAGAGCATTGACACACATAAAAAATTAAAACAACCCTTTGTACTTACGGTAGAA (/892<<
IL31_4368:1:1:998:19914 141 * 0 0 * * 0 0 GAATGAAAGCAGAGACCCTGATCGAGCCCCAGAAAGATACACCTCCAGATTTTA C?=CEC
```

SAM Example
Sorted SAM

One row is one read, NOT one fragment.

```
IL31_4368:1:107:15207:19097 163 chr1 17 0 54M = 21 58 CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
IL31_4368:1:107:15207:19097 83 chr1 21 0 54M = 17 -58 ACCCTACCCCTAGCCCTAACCCTACCCCTAACCCTAACCCTAACCCTAA

IL31_4368:1:10:17817:9758 137 chr1 23 0 54M = 23 0 CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCAGA

IL31_4368:1:54:13142:21400 163 chr1 37 0 54M = 44 61 TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
IL31_4368:1:54:13142:21400 83 chr1 44 0 54M = 37 -61 AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
```

SAM Specifications
Record Column

Col	Field	Type	Brief description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	Reference sequence NAME
4	POS	Int	1-based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33
12	META	metadata	

SAM Specifications
Record Column

Col	Field	Type
1	QNAME	IL31_4368:1:42:12530:7509
2	FLAG	137
3	RNAME	chr1
4	POS	10
5	MAPQ	30
6	CIGAR	54M
7	RNEXT	=
8	PNEXT	100
9	TLEN	90
10	SEQ	TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
11	QUAL	GGGGGGGFEGGGGCFGGGGGEGGFGEGGFGFGGFGFEGFC
12	META	XT:A:R NM:i:3 SM:i:0 AM:i:0 X0:i:11 X1:i:0 XM:i:3 XO:i:0 X

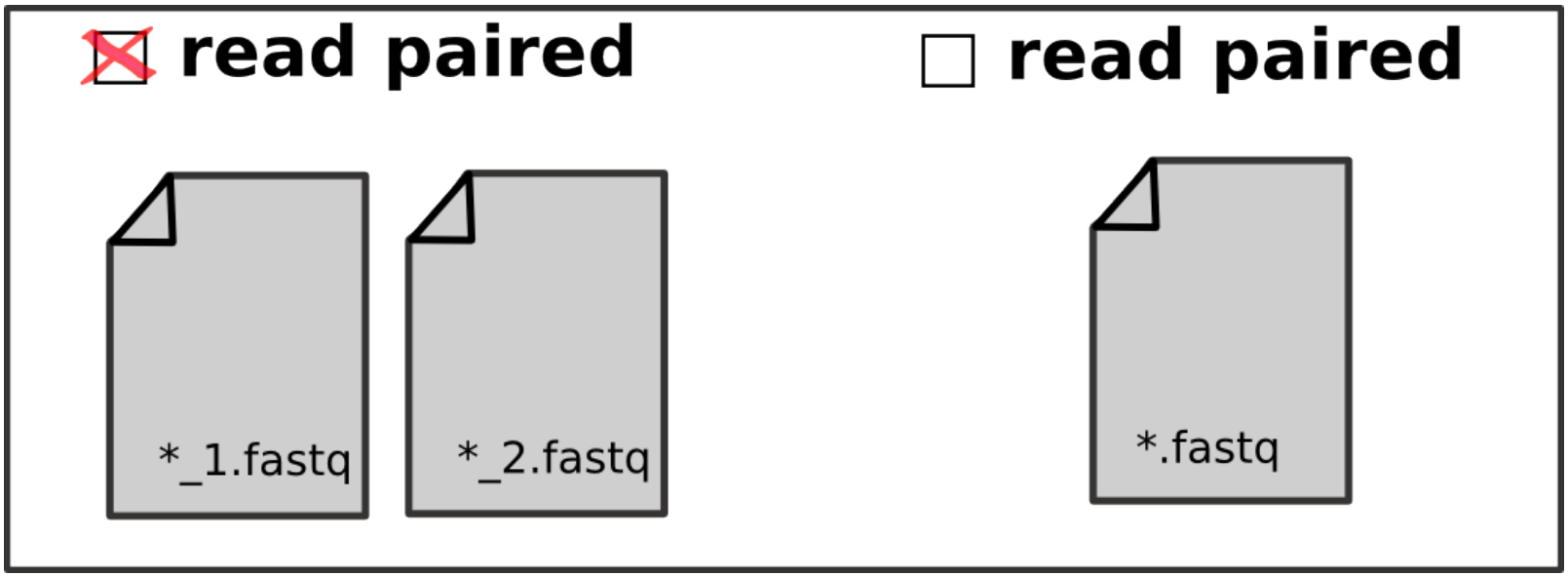
SAM FLAGS

- ☐ read paired.
- ☐ read mapped in proper pair.
- ☐ read unmapped.
- ☐ mate unmapped.
- ☐ read reverse strand.
- ☐ mate reverse strand.
- ☐ first in pair.
- ☐ second in pair.
- ☐ not primary alignment.
- ☐ read fails platform/vendor quality checks.
- ☐ read is PCR or optical duplicate

SAM FLAGS

BIT#:	7	6	5	4	3	2	1	0
	0	0	1	0	1	1	0	1
VALUE:	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
	128	64	32	16	8	4	2	1

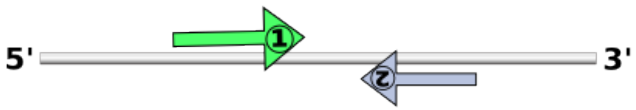
SAM FLAGS
Read Paired



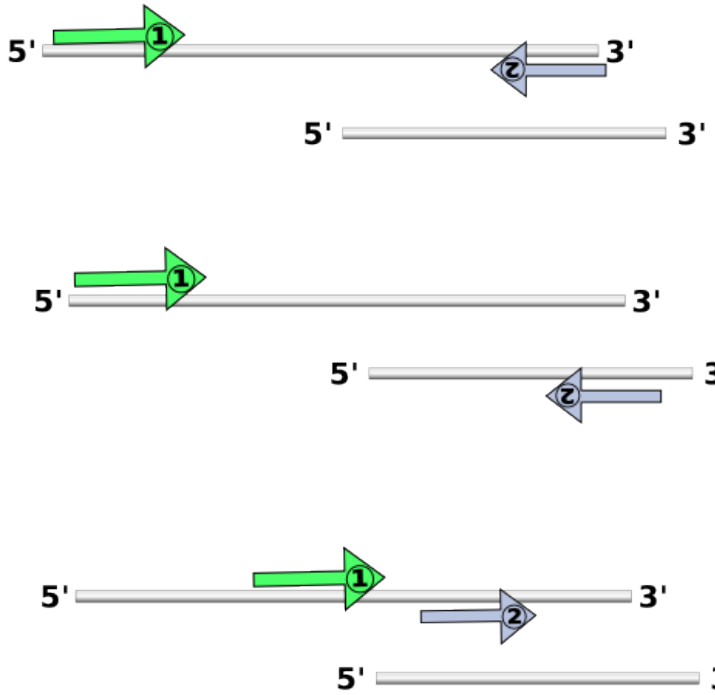
SAM FLAGS

Read mapped in proper pair

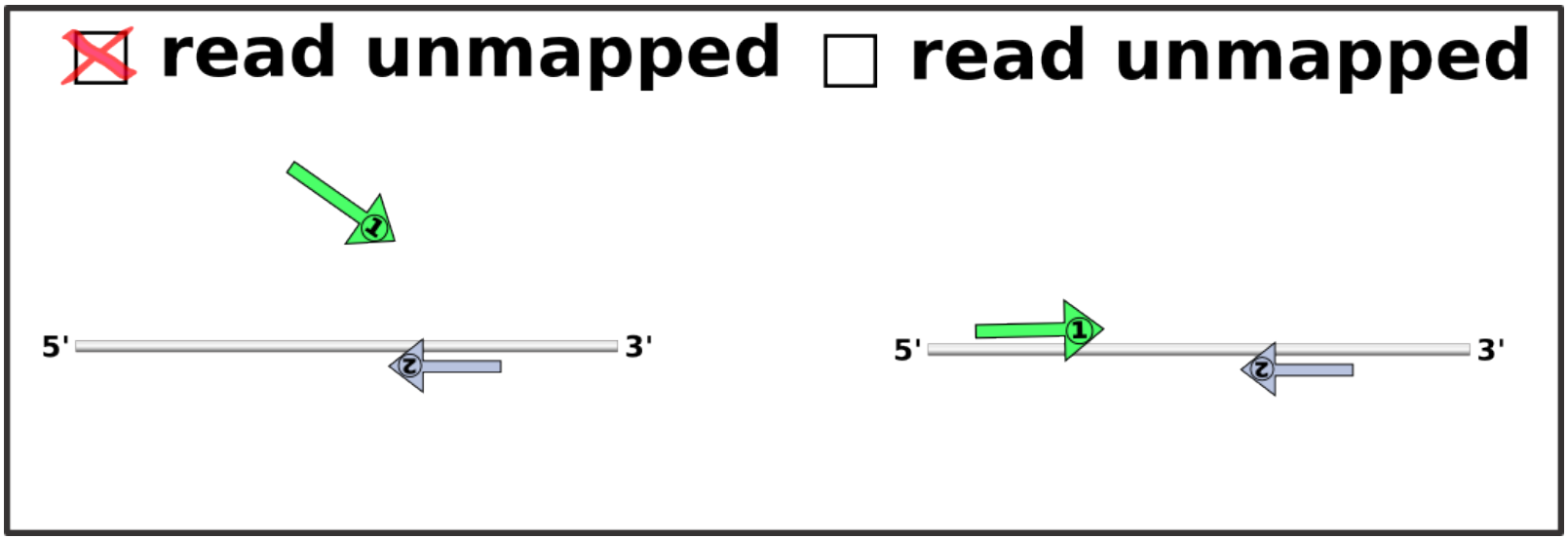
☒ **read mapped in proper pair**



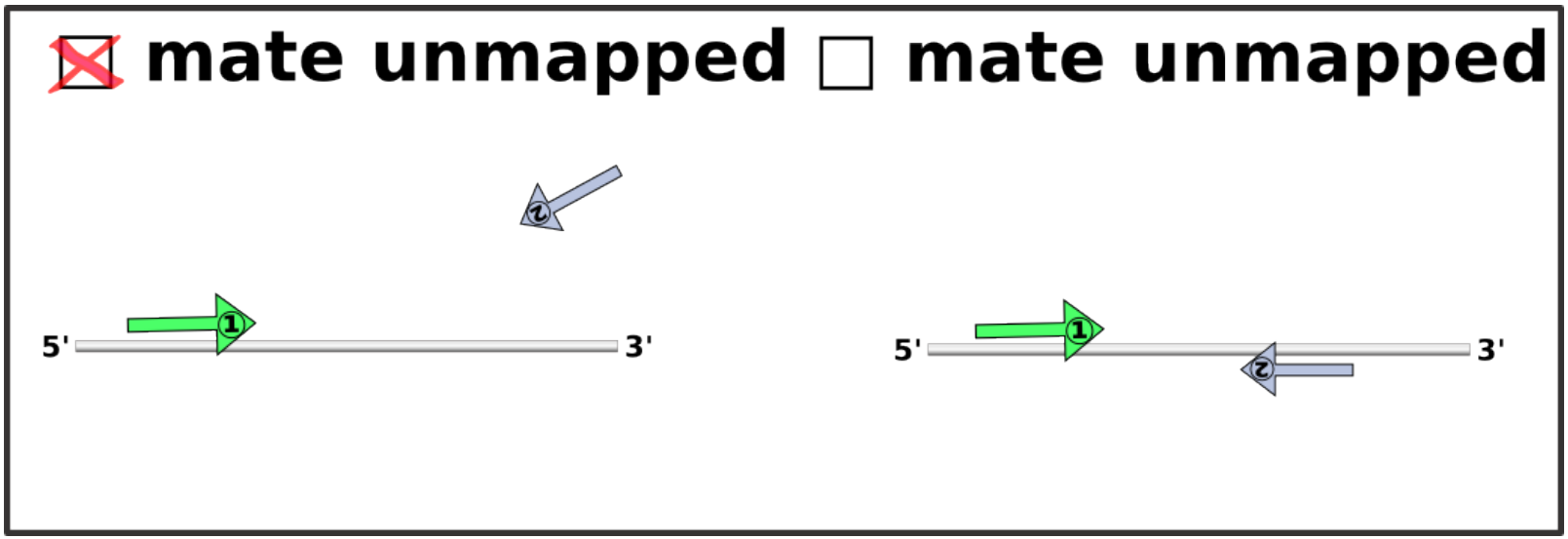
☐ **read mapped in proper pair**



SAM FLAGS
Read unmapped

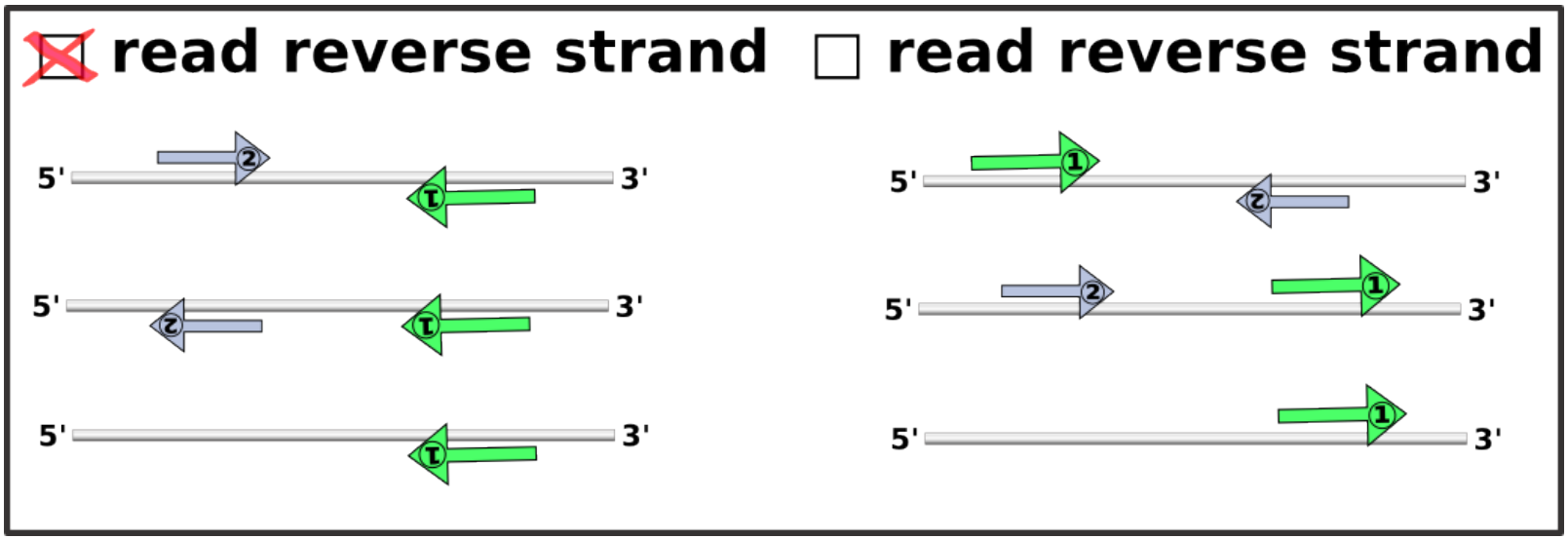


SAM FLAGS
Mate unmapped



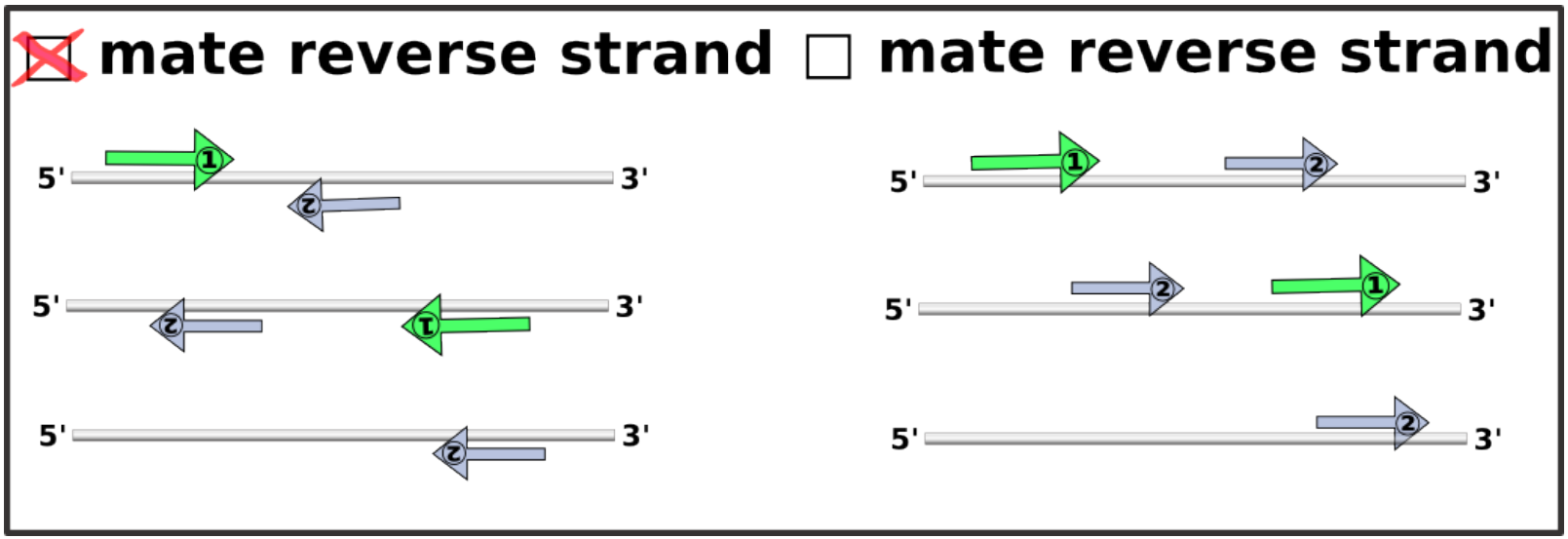
SAM FLAGS

Read reverse strand



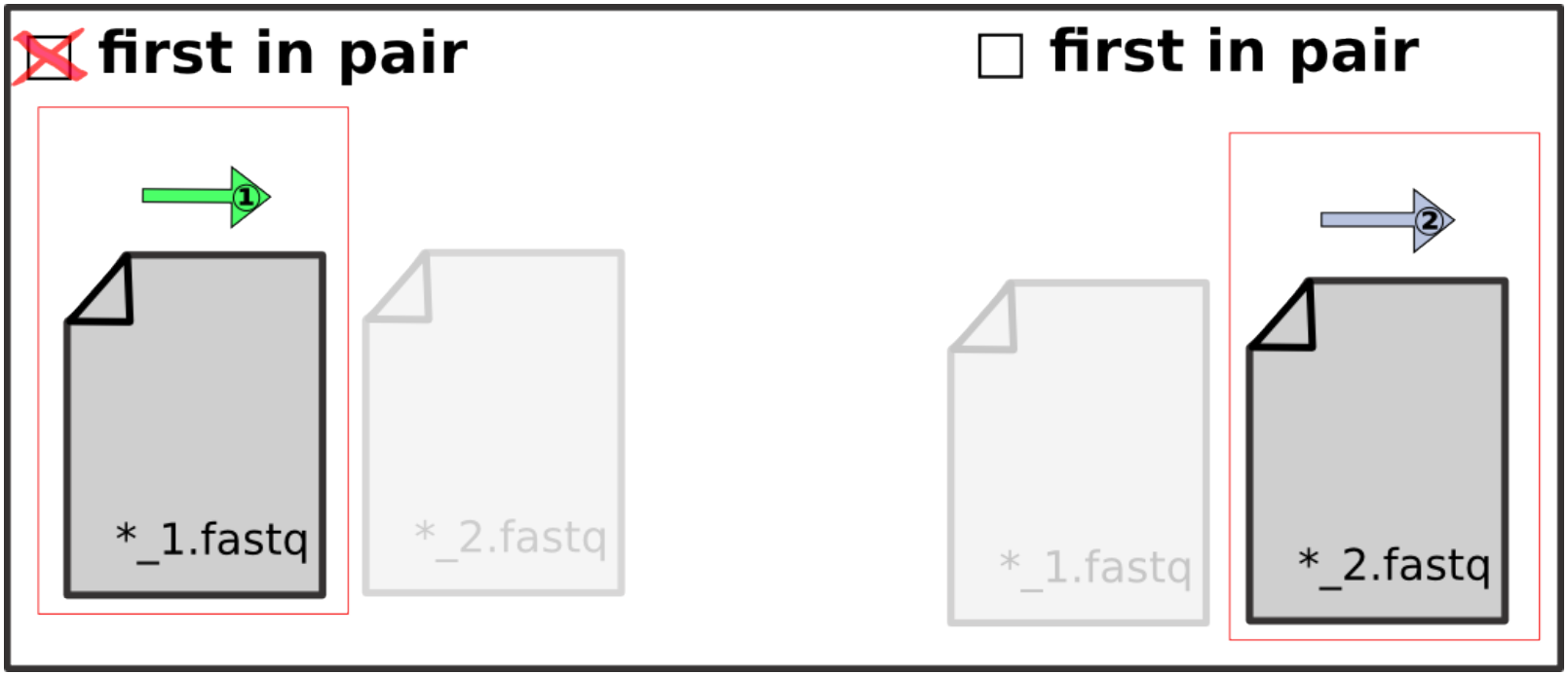
SAM FLAGS

Mate reverse strand

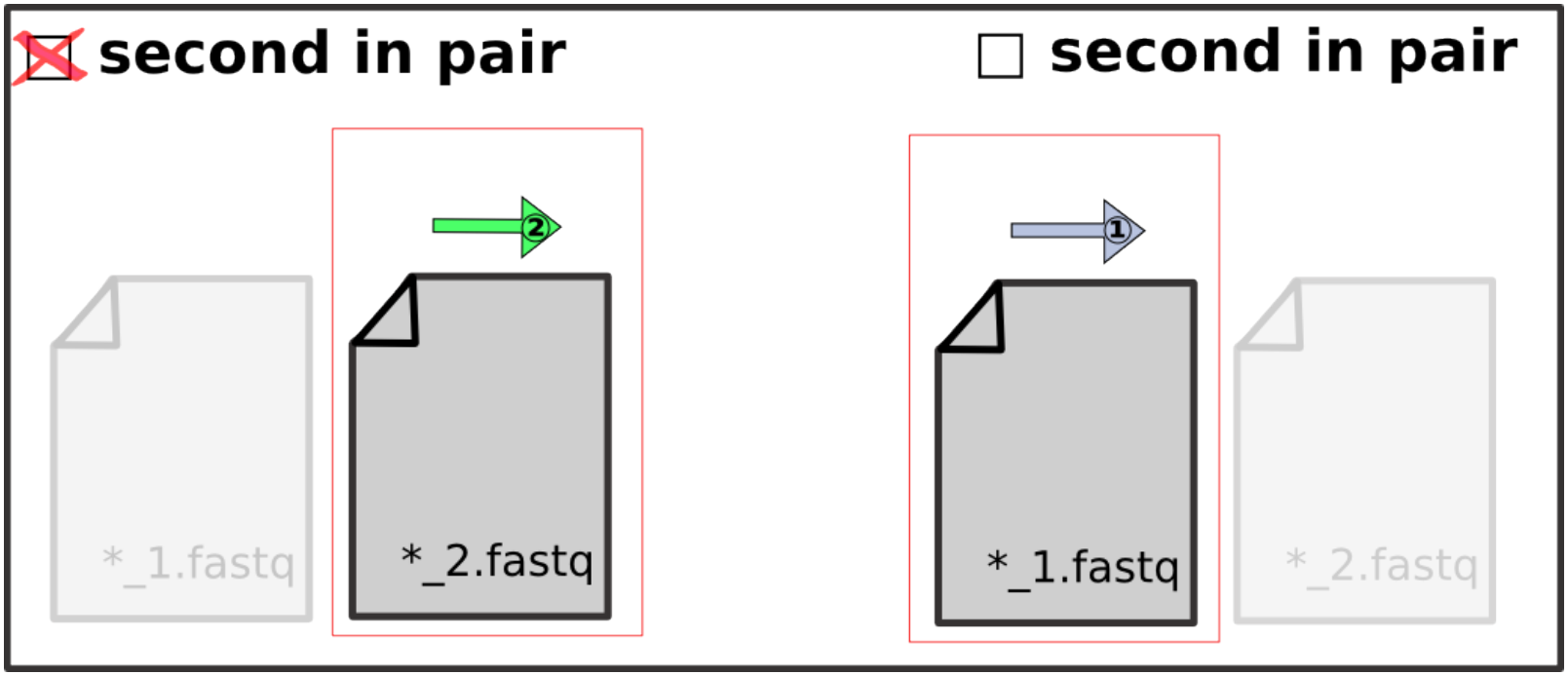


SAM FLAGS

First in pair

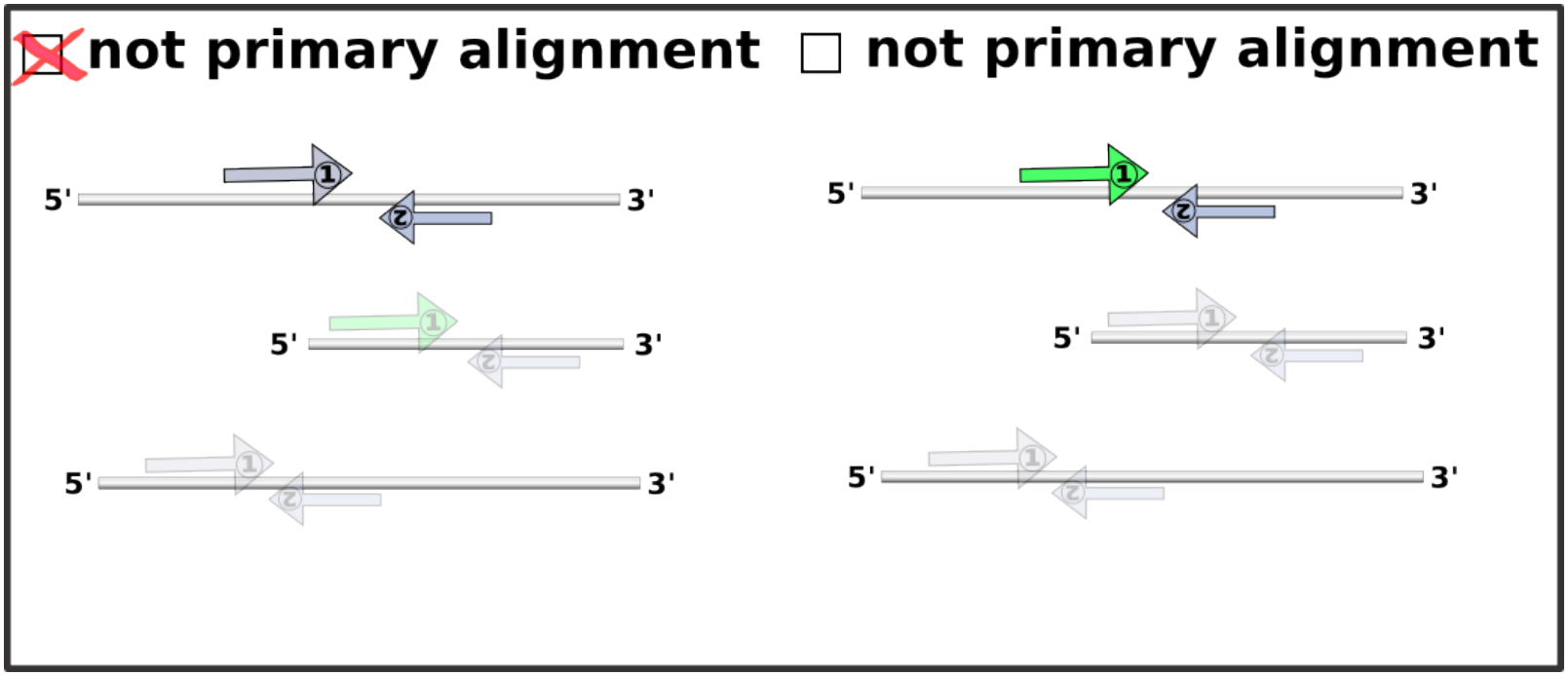


SAM FLAGS
Second in pair



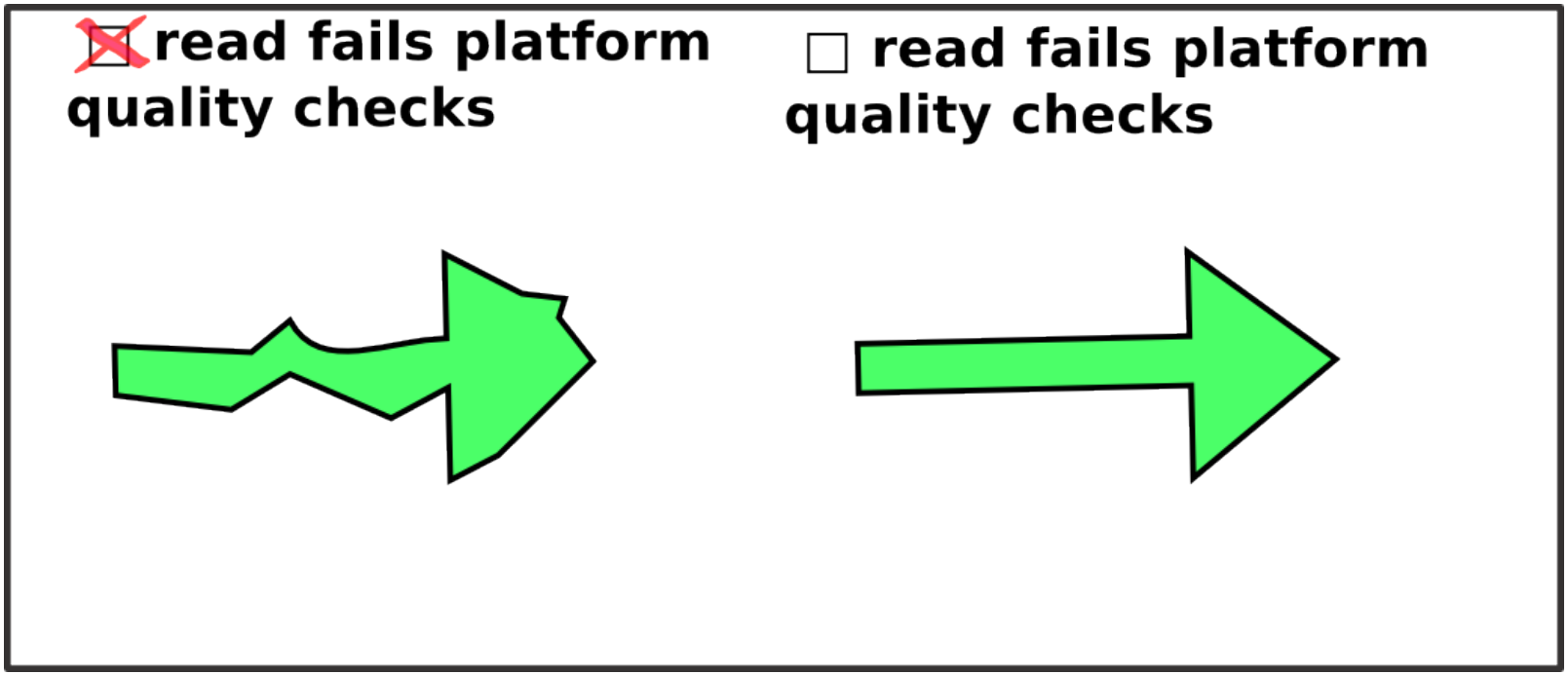
SAM FLAGS

not primary alignment



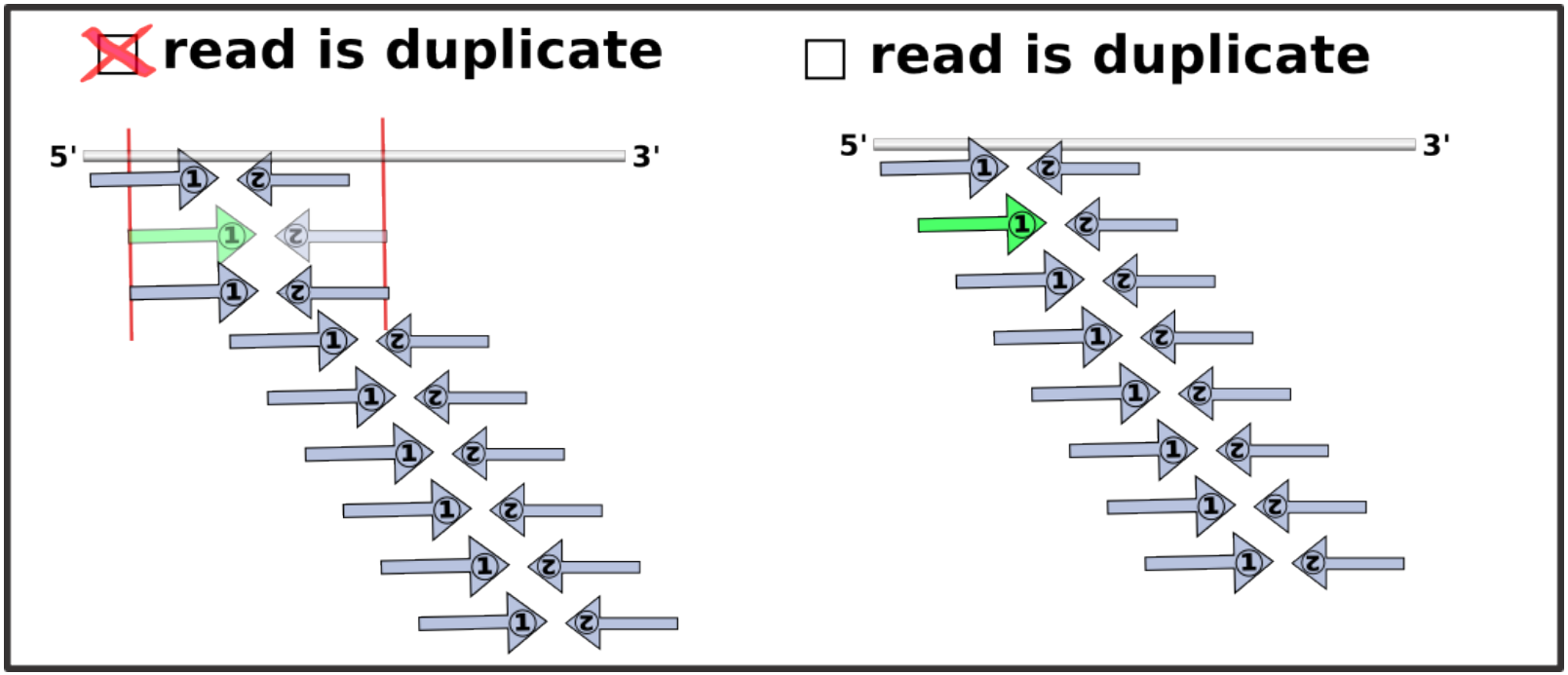
SAM FLAGS

read fails platform/vendor quality checks



SAM FLAGS

read is PCR or optical duplicate



SAM CIGAR

The CIGAR string is a sequence of of base lengths and the associated operation. They are used to indicate things like which bases align (either a match/mismatch) with the reference, are deleted from the reference, and are insertions that are not in the reference.

SAM Cigar

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

SAM Cigar

http://genome.sph.umich.edu/wiki/SAM

```
RefPos:     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

Aligning these two:

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

With the alignment above, you get:

POS: 5

CIGAR: 3M1I3M1D5M

or

CIGAR: 3=1I3=1D2=1X2=

optional fields on a SAM/BAM Alignment. A TAG is comprised of a two character TAG key, they type of the value, and the value:

`[A-Za-z][A-za-z]:[AifZH]:.*`

The types, A, i, f, Z, H are used to indicate the type of value stored in the tag.

Type	Description
A	character
i	signed 32-bit integer
f	single-precision float
Z	string
H	hex string

- XT:A:U - user defined tag called XT. It holds a character. The value associated with this tag is 'U'.
- NM:i:2 - predefined tag NM means: Edit distance to the reference (number of changes necessary to make this equal the reference, excluding clipping)

SAM Example
Sorted SAM

```
IL31_4368:1:107:15207:19097 163 chr1 17 0 54M = 21 58 CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
IL31_4368:1:107:15207:19097 83 chr1 21 0 54M = 17 -58 ACCCTACCCCTAGCCCTAACCCTACCCCTAACCCTAACCCTAACCCTAA

IL31_4368:1:54:13142:21400 163 chr1 37 0 54M = 44 61 TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
IL31_4368:1:54:13142:21400 83 chr1 44 0 54M = 37 -61 AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
```

BAM

BGZF Format

The SAM/BAM file format (Sequence Alignment/Map) comes in a plain text format (SAM), and a compressed binary format (BAM). The latter uses a modified form of gzip compression called BGZF (Blocked GNU Zip Format), which can be applied to any file format to provide compression with efficient random access

VCF
Variant Call Format

VCF Format

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome.

VCF
Example

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT     QUAL FILTER INFO
FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2
GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017
GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT
2/2:35:4
20     1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T
GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G
GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

- CHROM
- POS
- ID
- REF
- ALT
- QUAL
- FILTER
- INFO
- FORMAT
- SAMPLE-1
- SAMPLE-2
- SAMPLE-3
- ...

(...)

INFO fields should be described as follows

```
##INFO=<ID=ID,Number=number,Type=type,
    Description="description">
```

```
(...)
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
(...)
INFO                                    FORMAT      NA00001        NA00002        NA00003
20      14370   rs6054257 G       A       29    PASS    NS=3;DP=14;AF=0.5;DB;H2
GT:GQ:DP:HQ 0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20      17330   .         T       A       3     q10     NS=3;DP=11;AF=0.017
GT:GQ:DP:HQ 0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
```

FILTERs that have been applied to the data should be described as follows:

```
##FILTER=<ID=ID,Description="description">
```

```
(...)
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50 percent of samples have data">
(...)
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G       A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2
GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T       A       3    q10    NS=3;DP=11;AF=0.017
GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A       G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT
2/2:35:4
```

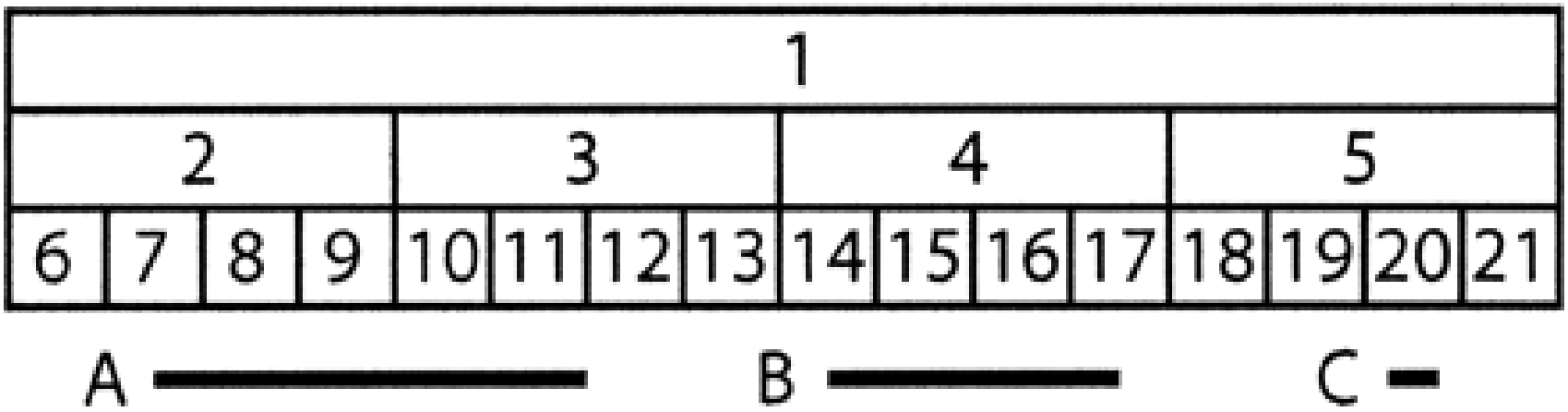
Genotype fields specified in the FORMAT field should be described as follows:

```
##FORMAT=<ID=ID,Number=number,Type=type,
    Description="description">
```

```
(...)
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
(...)
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1/0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A      3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0/1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP     2/2:35:4
20     1230237 .         T   .      47   PASS   NS=3;DP=13;AA=T                    GT:GQ:DP:HQ 0|0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Tabix

Binning



Building the TABIX index

```
$ bgzip -f  file.vcf
$ tabix -p vcf file.vcf.gz
```

Querying the TABIX index

```
$ tabix   file.vcf.gz chr3:1235-456778
```

API

Reading SAM with the samtools C library

```c
#include <stdlib.h>
#include <stdio.h>
#include "bam.h"
#include "sam.h"
int main(int argc, char *argv[]) {
   samfile_t* sam=samopen(argv[1], "rb", 0);
   bam1_t *b= bam_init1();
   long n=0L;
   while(samread(sam, b) > 0)
    {
    if(!(b->core.flag&BAM_FUNMAP)) ++n;
    }
   bam_destroy1(b);
   samclose(sam);
   printf("%lu\n",n);
   return 0;
   }
```

Reading SAM with the java picard library

```
import java.io.File;
import net.sf.samtools.*;
public class CountMapped {
  public static void main(String[] args) {
    long n=0L;
    File f=new File(args[0]);
    SAMFileReader sam = new SAMFileReader(f);
    for(SAMRecord rec : sam)
      {
      if(!rec.getReadUnmapped())
       {
       ++n;
       }
      }
    sam.close();
    System.out.println(n);
    }
```

End

Credits

- Angus: `http://ged.msu.edu/angus/`
- Wikipedia: `https://en.wikibooks.org/wiki/C%2B%2B_Programming/Programming_Languages/C%2B%2B/Code/Statements/Variables`
- Abecasis Group Wiki: `http://genome.sph.umich.edu/wiki/SAM`
- Genome Research `http://genome.cshlp.org/content/12/6/996`