INTRODUCTION AUX FORMATS DE FICHIERS

Plan

- 1. Formats de séquences brutes
  - 1.1. Format fasta
  - 1.2. Format fastq
- 2. Formats d’alignements
  - 2.1. Format SAM
  - 2.2. Format BAM
- 4. Format « Variant Calling »
  - 4.1. Format Varscan
  - 4.2. Format VCF

Format Fasta

```
>C10HBa0111D09_LR276    15142   24441   |Longueur=9300
GAACAAACAACCCCTTTTTGGAGGTGTTGGCGCGTCGTGCAGCTTACACTCAAAAGTTAA
AAAGTTGCCTTGCGATGCGGTCATGTTACAAACCTCTCTGCCTTAAATTAAATTCCATAA
CCAAGATTTGGAGGTGCCTCAACGATGCGCAGCCATGTCCCATATTTGGTCGCCTCGTTT
AAAAGTCAAGTTAGACTTAATTAAGAGGTCCAACTAGTGTAGGGGCGTTTTGAGTACTTG
TGGGATTTATTATAAACGGTTTTGAGTCACTTTAAACCCACTTCACCAATTAAAACAAAA
TCCTCAAGTTAAAACTCAATATCTTTCCATTCTCTCTCTCTAAAACCTTCATTGGAGATA
TTTGAAGCTCCACGGAAGAAGGTTAATTTTCCAAGGTTTCAATGAAAATTTCGTGTATAG
GTCTTCAATAAGGTATGGTGATTTCATCCTTGATTCTTCTATCATTCAAGGATCCAATTC
AAAGGTTTTTCAAAAGATCTCAAAAATCCTATTTCGAATTCTAAGTATGGGTTCTTCCAT
TTAAAGGTTTAAATGGATGAATTATGATGTTTTCAATGTTAGTTGATGTTTTTATGATAA
AAAAACTCCATGAACCCATGAGCATCCTAATTCTCTAATTTTGTCTTGTAAATTGAGTTT
GATAATTGTGATTGGTTATGGATGGAATTGTATTTAGATTGCTCTATATTGTTGATTCTT
ATTGTTAACCTATCTCTATATATGTAGAATTGAGATTGTAAGGATGAGTTAGTAATCTTG
GCTTTATGGGCTTTCGAATCCGGGTTTACCCCCTGGATGTAACCGGCATCCTCGCCCTTT
TTCAAGGACTAAGACCAACCTTTTAGTCTCATGTCATTACATTCATAGGTTGACAAATGC
GGAAAAATTTAAAACTTTCATTATCACTACTTGGAGGTTTACATAGACCTCTACATACAC
ATAAGATATATTCATATAGAGTATACATAGACCCTTCGTATAGGAAGGTTACATAGCCAT
CTACTTTTATTACACATACATATATATAAAATATAAAAATAGTCTAACGATTGTCTCATC
TCATACCCTCTAAACGATTATCACAATATGGGCATAACCCTTACATCAATCAAACAAGAG
CACATATAGGTCATACAAAAGTATAGTACTCAATTAAAAAGGAAAGAAATGAAAGAGTCT
TTAAGCTCATAACAAGTCCATAAGCTAGATTATGGCATTGACCTCAAAAGTTGAGGACCT
TATGTGCGTACACAAGCAAAACATGCTAAAAAGGGACTTTTTAGTCAAAACATGCCCATT
TATCCCTTTAAGAACCTACTACAAAGCCAACAAGTCATACCAACCAACCAAACATGCTTA
CTATCTCAACAAGTAATACTTATCCCAACATACTTGAAACCATGATTTACTACAACCCTA
TCACCAAGGAAAAATATCACAAGAATGAATAAGAGTCAATCATATCATGATAGAGAGACA
ACTATTCATGAATCCTTATCAACTCAACAAGTGCAATAACCAAGCAAAGCCTCATAACCT
TACTCAATCAAGTATCCTCAAAAAGAAACCATGACCAATGTCCAACTTTACCTAACATAG
CATTTAGGTTTACATTTTATCATATATTAACATTATGACCCAAGGCATACTCATTAGTAA
ACTAATTAATATATAATATCAACAATGTGCCATAGTAATCATATATACATAATATATCAT
```

Format fastq

- 1 séquence = 4 lignes dans le fichier

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- 1 ère ligne = identifiant de la séquence

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

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 number
ATCACG	index sequence

Format fastq

- 4ème ligne = Qualité

```
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- Appelée aussi Phred quality score (Sanger format)

$Q_{\text{sanger}} = -10 \log_{10} p$ Probabilité qu'une base soit incorrecte

Format fastq

- Encodée en ASCII (allège le fichier)

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Formats d’alignements

- Plusieurs format existent
  - SAM et BAM (= standards)
  - ELAND (spécifique Illumina)
  - MAQ Map

SAM Format: introduction

- NGS => a variety of new alignment tools : **Bowtie** (Langmead,B. et al (2009), **Maq** (Li,H. et al (2008), **BWA** (Li and Durbin, 2009), ...
- SAM : a common alignment format that supports all sequence types and aligners
- SAM : **S**equence **A**lignment/**M**ap format
- A well-defined interface between alignment and downstream analyses

overview

@SQ SN:C09SLm0143I09_LR365 LN:10488

@PG ID:Bowtie VN:0.12.7 CL:"bowtie -q -X 1000 --fr -p 4 -S --phred33-quals /tmp/2008773.1.galaxy.q/tmp646AgK/tmprUje1z -1 /galaxy/galaxy-dist/database/files/001/dataset_1812.dat -2 /galaxy/galaxy-dist/database/files/001/dataset_1813.dat"

HWI-EAS337_3:7:1:415:1217	163	C02HBa0185P07_LR40	3830	255	36M	=	3889	95	TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC	78887877777776787877887877765555/	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:415:1217	83	C02HBa0185P07_LR40	3889	255	36M	=	3830	-95	ACAGTGATGTAGTCCTGCGTGAAAAGTCTGCACATC	256266878778817,77778888788818777888	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:1178:755	163	C11SLe0053P22_LR298	1980	255	36M	=	2130	186	GACATTTCAATTACATTCATCTTACCATCACCTATA	878788888788788788778887787877555553	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:1178:755	83	C11SLe0053P22_LR298	2130	255	36M	=	1980	-186	ATTCAATGGTTTTACCATCAACCAACCACTCTCACC	66666677787877778778877888788888888	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:278:1153	77	*	0	0	*	*	0	0	GAGAAAACCTGTAATAAATACTGAGAGAAAGTAGGG	88888888888888888888887877887666663	XM:i:0
HWI-EAS337_3:7:1:278:1153	141	*	0	0	*	*	0	0	GTCAGGCCGCATTGATGGGGGATGGGTTTCCCCCCA	88878888888887777777777888777555553	XM:i:0
HWI-EAS337_3:7:1:208:1489	77	*	0	0	*	*	0	0	GGAAACATATGCACATAAACGTTGAAATCATGCTTA	88888888888888888788788788888666666	XM:i:0
HWI-EAS337_3:7:1:208:1489	141	*	0	0	*	*	0	0	CGTGTTTTTGGTTGTGCATAAGGCTTTTTAAAGTAA	88888888877882878788767888888735355 3	XM:i:0
HWI-EAS337_3:7:1:277:1259	99	C06HBa0144J05_LR355	1	255	36M	=	101	136	GGGTGACAAAGAAAACAAAAGGGACATGGTACTTGG	88888888888888888888887888888 7666666	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:277:1259	147	C06HBa0144J05_LR355	101	255	36M	=	1	-136	TCTTCAAGTGATTCAGAAGATCCTGATGAGCCAAAA	455535887887777888887788888888888888	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:1154:1517	163	C02HBa0329G05_LR52	4680	255	36M	=	4746	102	CTAACTTCAATAATCAAGCTTGTCAGTGGAAGAAAA	88888888778778877788787777777555553	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:1154:1517	83	C02HBa0329G05_LR52	4746	255	36M	=	4680	-102	TGTGCTTCATAGGTAGGAGTAAGGTCTGCAACATTC	66664687878788888888888878878888888	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:447:1231	163	C08HBa0165B06_LR218	3575	255	36M	=	3619	80	TCAACAAGAGAAAGGAGACGAAAAAGTAAATCCAAC	88888888788887788888778888778555553	XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:447:1231	83	C08HBa0165B06_LR218	3619	255	36M	=	3575	-80	AGGCTCCAGCTTTCCATTCCAACTCTTCCACAAGTC	66463677777777788788787878888888888	XA:i:0 MD:Z:36 NM:i:0

@SQ name LN :

ref seq name

ref seq length

header

readName1 flag referenceName

readName2 flag referenceName

...

alignment

```
@SQ SN:C09SLm0094A22_LR246 LN:10193
@SQ SN:C09SLm0129J22_LR373 LN:9064
@SQ SN:C09SLm0143I09_LR365 LN:10488
@PG ID:Bowtie VN:0.12.7 CL:"bowtie -q -X 1000 --fr -p 4 -S --phred33-quals /tmp/2008773.1.galaxy.q/tmp646AgK/tmprUje1z -1 /galaxy/
galaxy-dist/database/files/001/dataset_1812.dat -2 /galaxy/galaxy-dist/database/files/001/dataset_1813.dat"
HWI-EAS337_3:7:1:415:1217       163     C02HBa0185P07_LR40      3830    255     36M     =       3889    95      TAAGAACTTGGCTGATCGCCTA
CTTACTGCTTTTAC  78887877777777678787788787776755555/    XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:415:1217       83      C02HBa0185P07_LR40      3889    255     36M     =       3830    -95     ACAGTGATGTAGTCCTGCGTGA
AAAGTCTGCACATC  256266878778817,77778888788818777888    XA:i:0 MD:Z:36 NM:i:0
HWI-EAS337_3:7:1:1178:755       163     C11SLe0053P22_LR298     1980    255     36M     =       2130    186     GACATTTCAATTACATTCATCT
```

one tab-delimited line per alignment

Tab-delimited : SAM Fields

Table 1. Mandatory fields in the SAM format

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

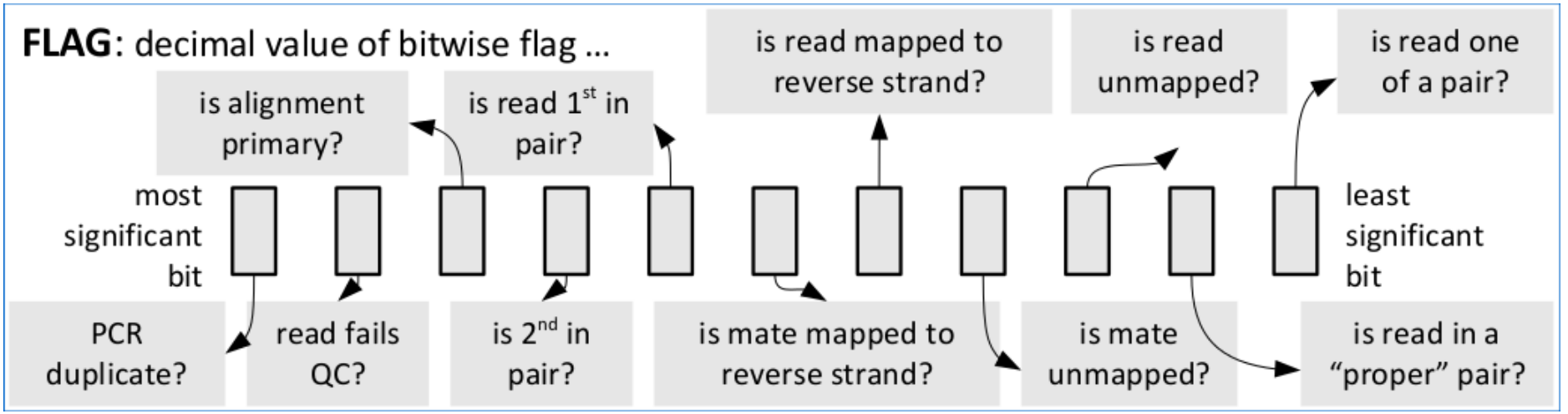
SAM Format (example)

```
HWI-EAS337_3:7:1:415:1217       163     C02HBa0185P07_LR40      3830    255     36M     =       3889    95      TAAGAACTTGGCTGATCGCCTA
CTTACTGCTTTTAC  78887877777777678787788787776755555/    XA:i:0 MD:Z:36 NM:i:0
```

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777777678787788787776755555
XA:i:0 MD:Z:36 NM:i:0

QNAME: Query name

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0



HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

163 (decimal) = 00010100011(binairy)
-read is one of a pair
-each segment properly aligned according to the aligner
-read is in second pair
-read n°1 is mapped on reverse strand

http://picard.sourceforge.net/explain-flags.html

RNAME : reference sequence name

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

POS : position on reference

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777777678787788787776755555
XA:i:0 MD:Z:36 NM:i:0

MAPQ : mapping quality

It equals **−10 log10 Pr{mapping position is wrong}**, rounded to the nearest integer.

A value 255 indicates that the mapping quality is not available.

Zero value is the lowest quality.

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

CIGAR : episode 1

CIGAR : extended CIGAR string (Compact Idiosyncratic Gapped Alignment Report)
Format: [0-9][MIDNSHP][0-9][MIDNSHP]...
[0-9] : position
M = match or mismatch (?!), I/D = insertion / deletion, N = skipped bases on reference, S/H = soft / hard clip (soft means nt's still appear in sequence field), P = padding
e.g.: ”1S81M” means that the first (5'-most) nt is not part of the alignment, but the following 81 nt's are either matches or mis-matches.

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

RNEXT : mate or not mate ?

' = ' means the mate is mapped to the same reference sequence as the current read

' * ' means that the read is unpaired (has no mate)

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

PNEXT : mate position

' 0 ' means no info is avalaible

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

TLEN : insert size

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

SEQ and **QUAL** : sequence and quality c.f. FASTQ format

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877676755555
XA:i:0 MD:Z:36 NM:i:0

OPT : optional fields
Follow the TAG:TYPE:VALUE format.
TYPE is : [A(printable character); i(signed integer); f(floating point); z(printable string);
H(hex string)]
Ex : NM:i:0 edit distance, equal zero for this alignment

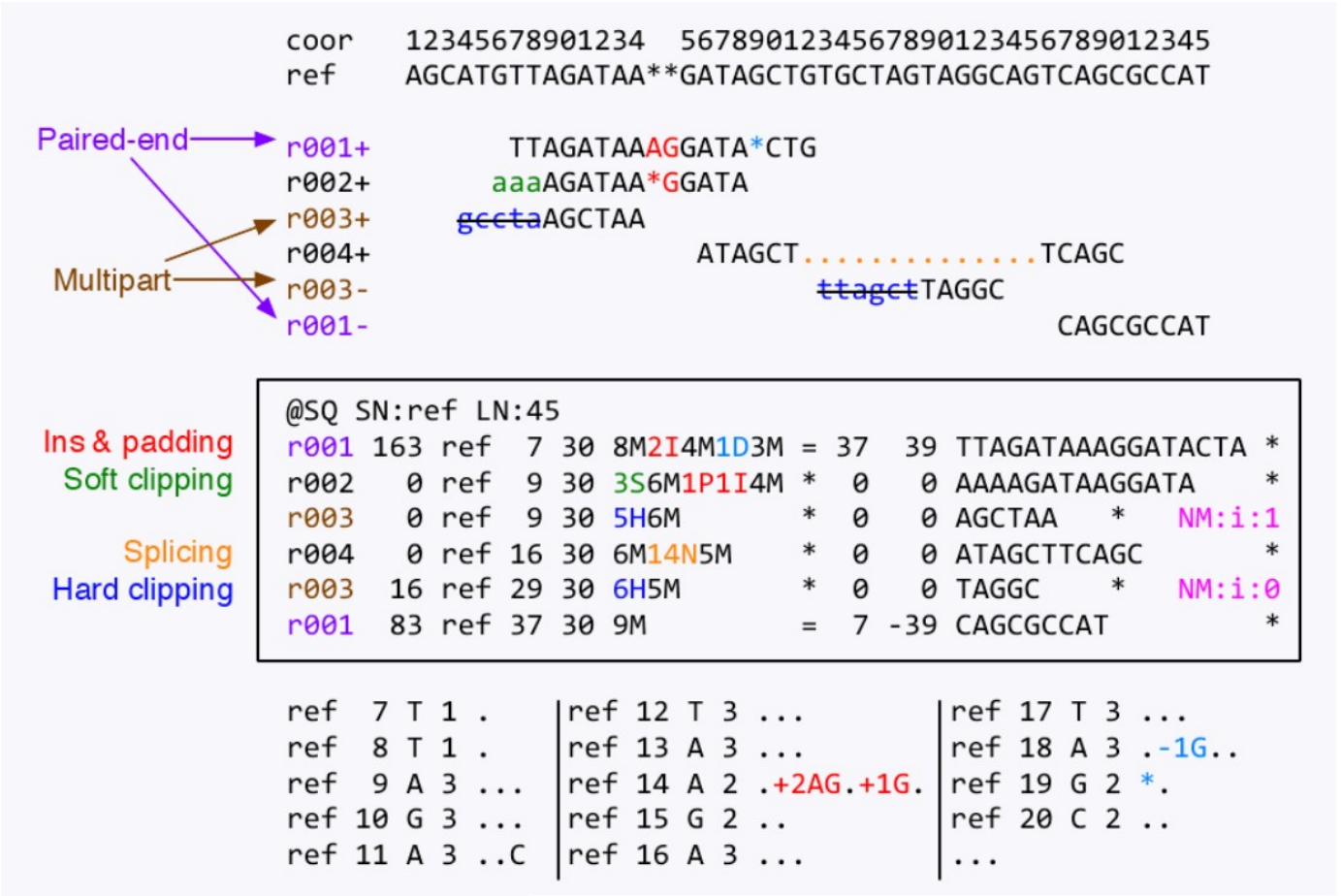
HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

CIGAR : episode 2

CIGAR : extended CIGAR string (Compact Idiosyncratic Gapped Alignment Report)
Format: [0-9][MIDNSHP][0-9][MIDNSHP]...
[0-9] : position
M = match or mismatch (?!), I/D = insertion / deletion, N = skipped bases on reference, S/H = soft / hard clip (soft means nt's still appear in sequence field), P = padding
e.g.: ”1S81M” means that the first (5'-most) nt is not part of the alignment, but the following 81 nt's are either matches or mis-matches.

HWI-EAS337_3:7:1:415:1217
163
C02HBa0185P07_LR40
3830
255
36M
=
3889
95
TAAGAACTTGGCTGATCGCCTACTTACTGCTTTTAC
78887877777776787877887877776755555
XA:i:0 MD:Z:36 NM:i:0

CIGAR : episode 2



BAM

- BAM = compressed SAM
- Indexed BAM : *.bam.bai
- Tools (post process, viewers) use indexed bam to avoid all information extraction

SAM Tools

SAM Tools : a library and software package for parsing and manipulating alignments in the SAM/BAM format.

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format

Picard Tools

NGS: Picard (beta)

CONVERSION

- SAM to FASTQ creates a FASTQ file

QC/METRICS FOR SAM/BAM

- SAM/BAM Alignment Summary Metrics
- SAM/BAM GC Bias Metrics
- SAM/BAM Hybrid Selection Metrics for targeted resequencing data

BAM/SAM CLEANING

- Reorder SAM/BAM
- Replace SAM/BAM Header

Reference

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

Format VARSCAN 2.2

Chrom	chromosome name
Position	position (1-based)
Ref	reference allele at this position
Var	variant allele at this position
Reads1	reads supporting reference allele
Reads2	reads supporting variant allele
VarFreq	frequency of variant allele by read count
Strands1	strands on which reference allele was observed
Strands2	strands on which variant allele was observed
Qual1	average base quality of reference-supporting read bases
Qual2	average base quality of variant-supporting read bases
Pvalue	Significance of variant read count vs. expected baseline error

Varscan 2.2 Example

Chrom	Position	Ref	Var	Reads1	Reads2	VarFreq	Strands1	Strands2	Qual1	Qual2	Pvalue
chr1	1252920	A	G	23	7734	99.7%	2	2	60	64	0.0
chr1	1252999	A	G	20	7785	99.74%	2	2	60	63	0.0
chr1	1253107	T	A	13	3538	99.63%	2	2	64	63	0.0
chr1	3516323	A	G	17	1327	98.74%	2	2	61	63	0.0
chr1	3516329	T	C	15	1530	99.03%	2	2	62	64	0.0
chr1	3516333	T	C	20	1975	99%	2	2	59	64	0.0
chr1	3516335	T	C	16	2252	99.29%	2	2	62	64	0.0

Format VARSCAN 2.2.8

Chrom	chromosome name
Position	position (1-based)
Ref	reference allele at this position
Cons	Consensus genotype of sample in IUPAC format.
Reads1	reads supporting reference allele
Reads2	reads supporting variant allele
VarFreq	frequency of variant allele by read count
Strands1	strands on which reference allele was observed
Strands2	strands on which variant allele was observed
Qual1	average base quality of reference-supporting read bases
Qual2	average base quality of variant-supporting read bases
Pvalue	Significance of variant read count vs. expected baseline error
MapQual1	Average map quality of ref reads (only useful if in pileup)
MapQual2	Average map quality of var reads (only useful if in pileup)
Reads1Plus	Number of reference-supporting reads on + strand
Reads1Minus	Number of reference-supporting reads on - strand
Reads2Plus	Number of variant-supporting reads on + strand
Reads2Minus	Number of variant-supporting reads on - strand
VarAllele	Most frequent non-reference allele observed

VARSCAN 2.2.8 Example

Chrom	Pos.	Ref	Cons	R1	R2	VarFreq	Str1	Str2	Q1	Q2	Pval	MapQ1	MapQ2	R1+	R1-	R2+	R2-	VarAllele
C12HBa115G22_LR301	1198	A	R	1	1	50%	1	1	23	23	0.98	1	1	1	0	1	0	G
C02HBa0072A04_LR26	10	G	K	1	1	50%	1	1	22	23	0.98	1	1	1	0	1	0	T
C02SLe0018B07_LR335	8941	C	T	0	2	100%	0	1	0	20	0.98	0	1	0	0	0	2	T
C05HBa0145P19_LR136	2824	G	A	0	2	100%	0	1	0	22	0.98	0	1	0	0	2	0	A
C06HBa0217M17_LR166	660	C	M	1	1	50%	1	1	23	23	0.98	1	1	0	1	0	1	A
C07HBa0018L21_LR201	8890	A	R	1	1	50%	1	1	22	22	0.98	1	1	1	0	1	0	G

Format VCF

- SAM = standard for alignment
- VCF = standard for storing sequence variation
- SNPs, indels, large structural variants
- Primary intention : to represent human genetic variation (1000 Genome Project)
- Can be used in diffrent contexts

overview

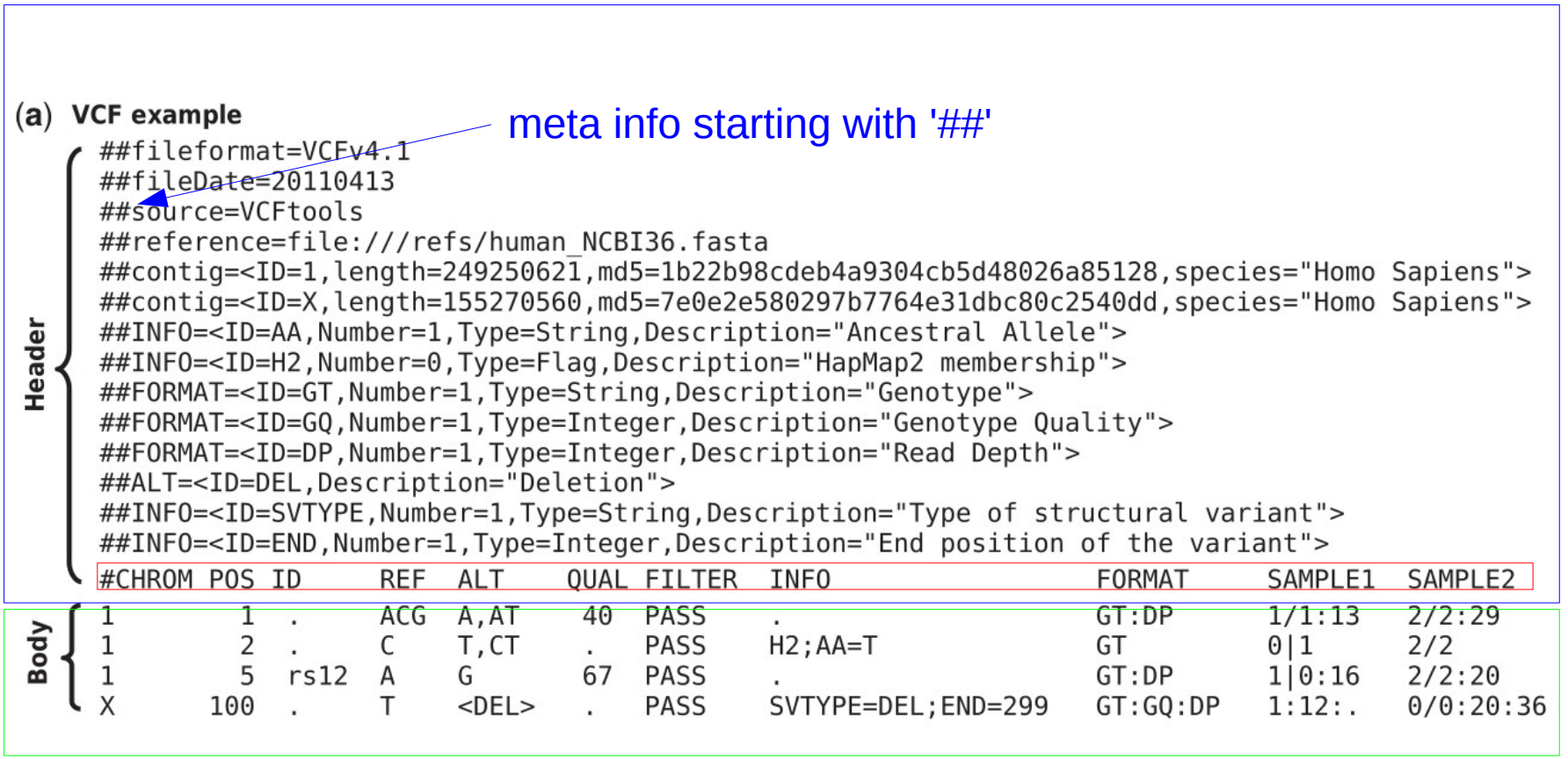
(a) VCF example

Header

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0|1	2/2
1	5	rs12	A	G	67	PASS	.	GT:DP	1|0:16	2/2:20
X	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36



Meta info : provide a standardized description of tags and annotations used in the body section.

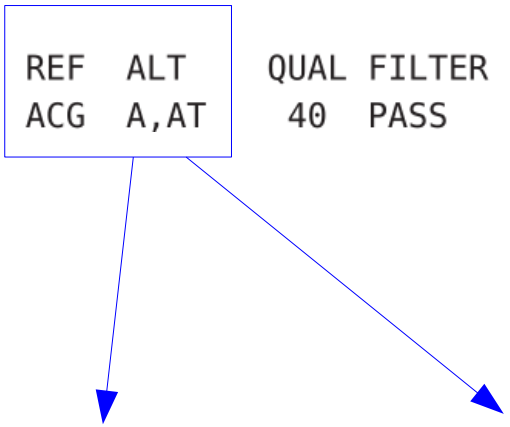
example

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29

mandatory fields

#CHROM : chrom id
#ID : unique identifier of variant
#POS : position of the start of the variant
#REF : refrence allele
#ALT : comma seprated list of alternate non reference alleles
#QUAL : phred quality score (?)
#FILTER : site filtering information (?)
#INFO : user extensible annotation

```
#CHROM POS ID      REF  ALT    QUAL FILTER  INFO                FORMAT      SAMPLE1  SAMPLE2
1        1  .      ACG  A,AT    40  PASS    .                   GT:DP       1/1:13   2/2:29
```



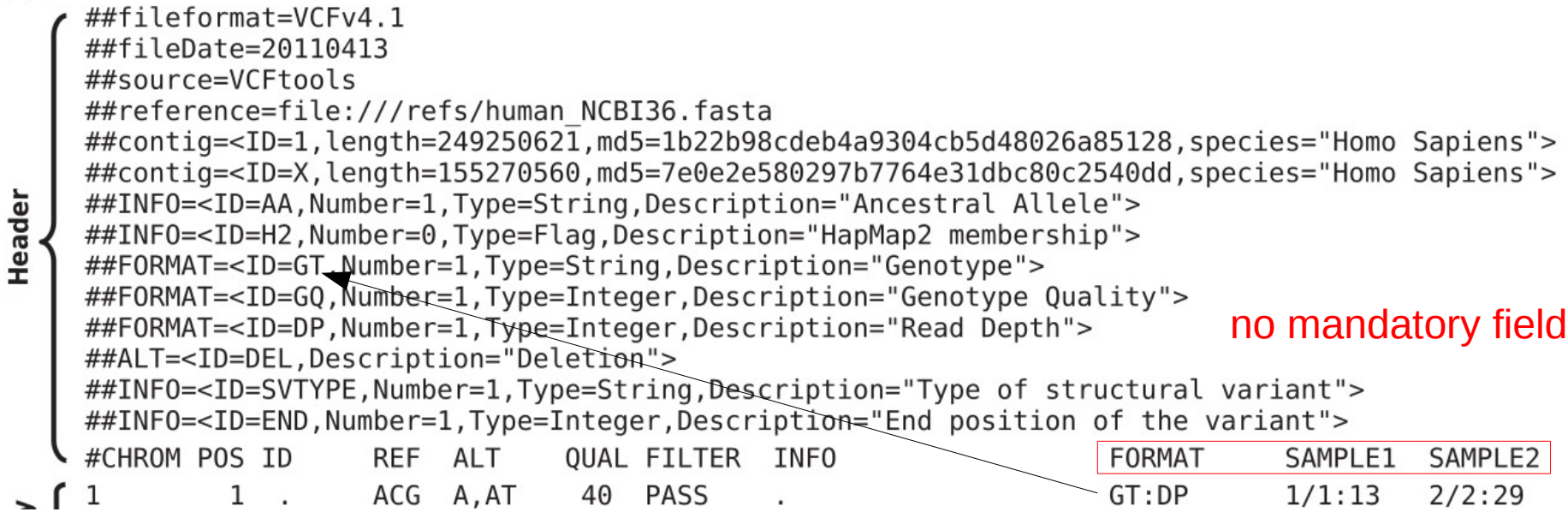
(d) Deletion

```
1234   POS REF ALT
ACGT   1   ACG A
A--T
 ^^
```

(e) Replacement

```
1234    POS REF ALT
ACGT    1   ACG AT
A-TT
 ^^
```

(a) **VCF example**



#FORMAT : describe format of #SAMPLE(s)
#FORMAT : infos found in the header
Samples for this line : genotypes and read depth

#SAMPLE1: genotype '1 ' (i.e. deletion) is on each allele and read depth is 13

#SAMPLE 2: genotype '2 ' (i.e remplacement) is on each allele and read depth is 29

example (2)

```
#CHROM POS ID    REF  ALT   QUAL FILTER INFO              FORMAT     SAMPLE1  SAMPLE2
1      1   .     ACG  A,AT   40  PASS   .                 GT:DP      1/1:13   2/2:29
1      2   .     C    T,CT    .  PASS   H2;AA=T           GT         0|1      2/2
```

(b) SNP

Alignment

```
1234
ACGT
ATGT
 ^
```

VCF representation

```
POS REF ALT
2   C   T
```

(c) Insertion

```
12345
AC-GT
ACTGT
  ^
```

```
POS REF ALT
2   C   CT
```

(a) **VCF example**

Header

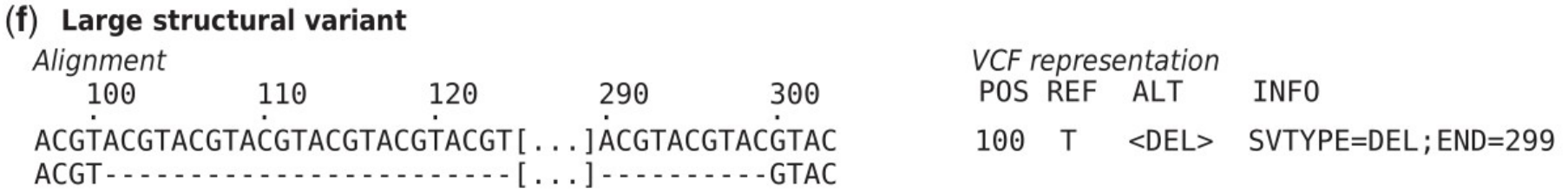
```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                FORMAT     SAMPLE1  SAMPLE2
1     1   .      ACG  A,AT    40   PASS    .                   GT:DP      1/1:13   2/2:29
1     2   .      C    T,CT    .    PASS    H2;AA=T             GT         0|1      2/2
```

#INFO: found in the header
Ancestral allele is 'T', variants found is HapMap2 membership

#FORMAT
Samples for this line : genotype

#SAMPLE1 : one allele with genotype '0' (0 is reference) and one allele with genotype '1 ' (SNP)
#SAMPLE2 : genotype '2' (insertion) on each allele

large struct variant



VCF Tools

- VCF Tools = 2 modules
- Operations on VCF files : format validation, merging, comparing, intersecting
- Analyse SNP data in VCF format : allele frequencies, various Quality Control metrics
- GATK toolkit : alternative tools for VCF generation and manipulation

Reference

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 27 no. 15 2011, pages 2156–2158*
doi:10.1093/bioinformatics/btr330

Sequence analysis

Advance Access publication June 7, 2011

The variant call format and VCFtools

Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

Sources & Ref

- Joe Fass <jnfass@ucdavis.edu> and his « Next Generation Sequence Alignment » slides
- The Sequence Alignment/Map format and SAM tools. Li *et al.* 2009 Bioinformatics 25 2078-2079
- The variant call format and VCFtools. Daneck *et al.* 2011 Bioinformatics 27 2156-2518.