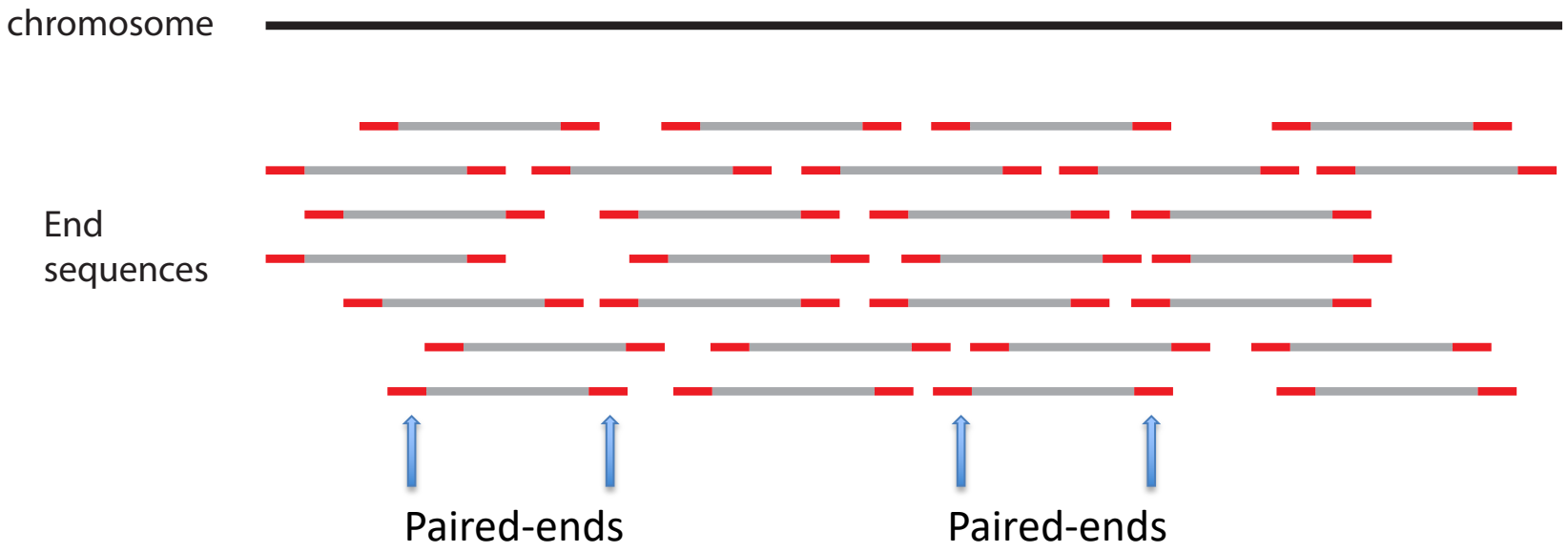
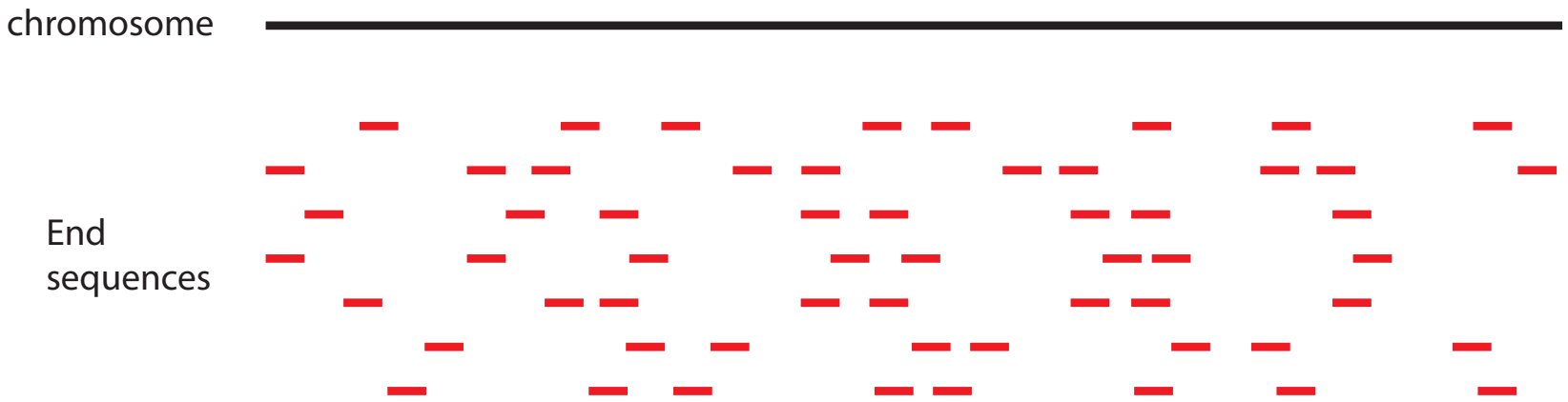
Short read assembly

Shear genomic DNA

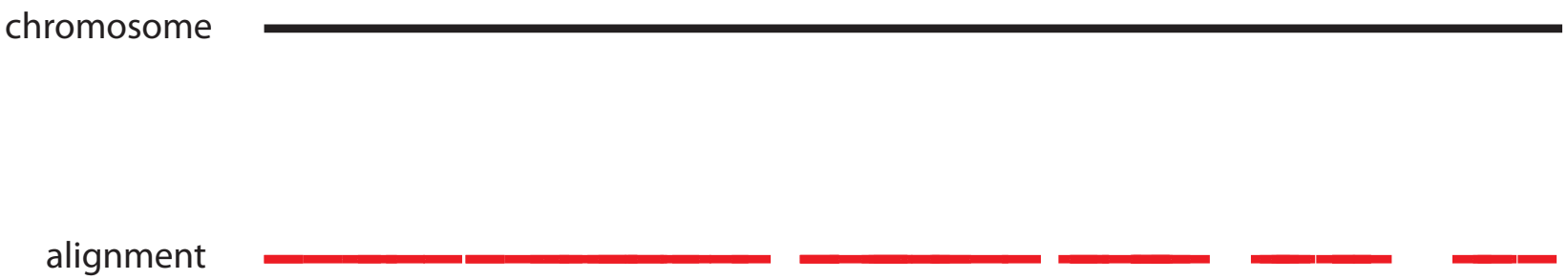
Sequence both ends of each fragment



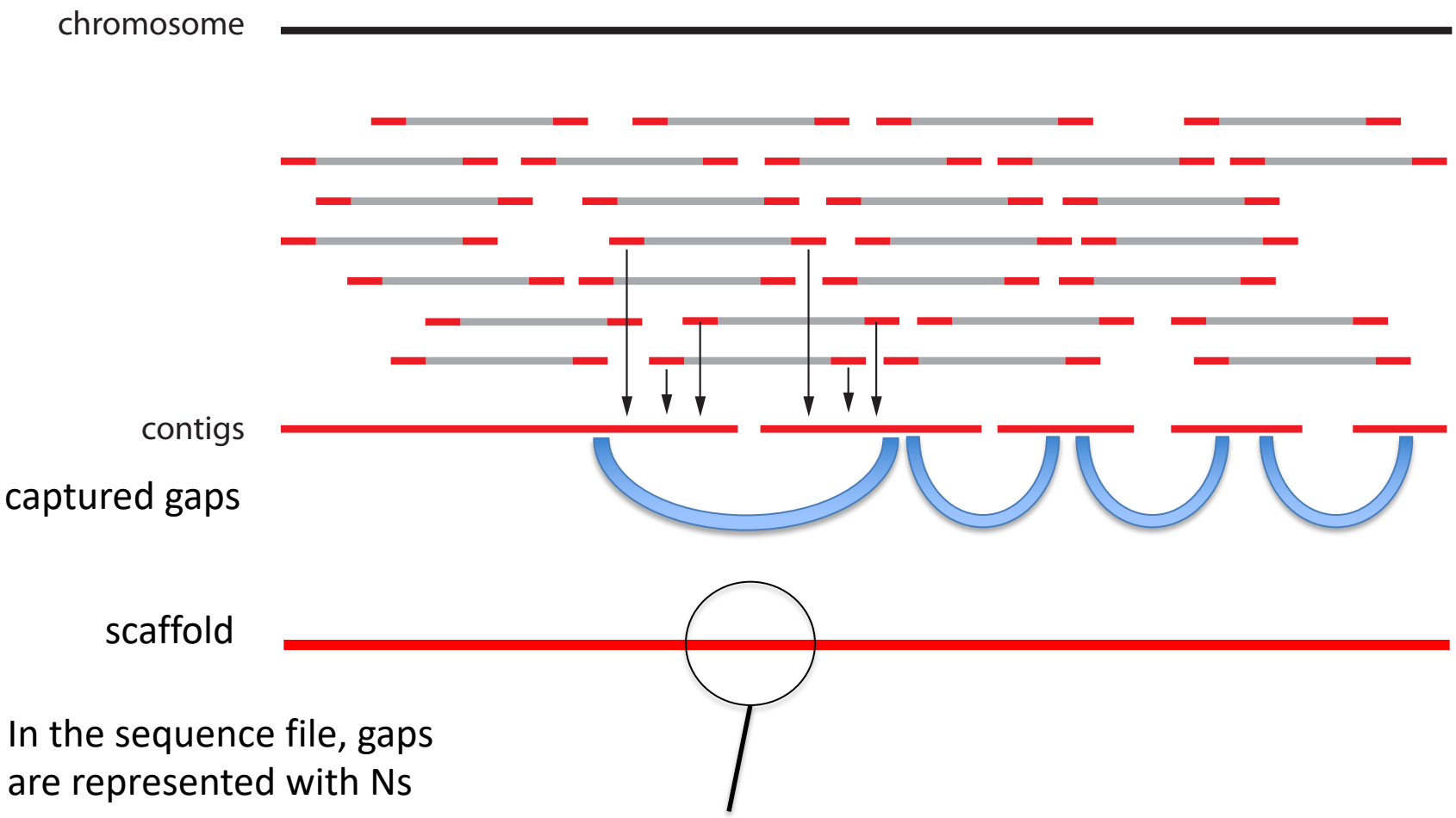
Sequence both ends of each fragment



Align sequence reads to form contigs



Paired ends allow linking of contigs into scaffolds

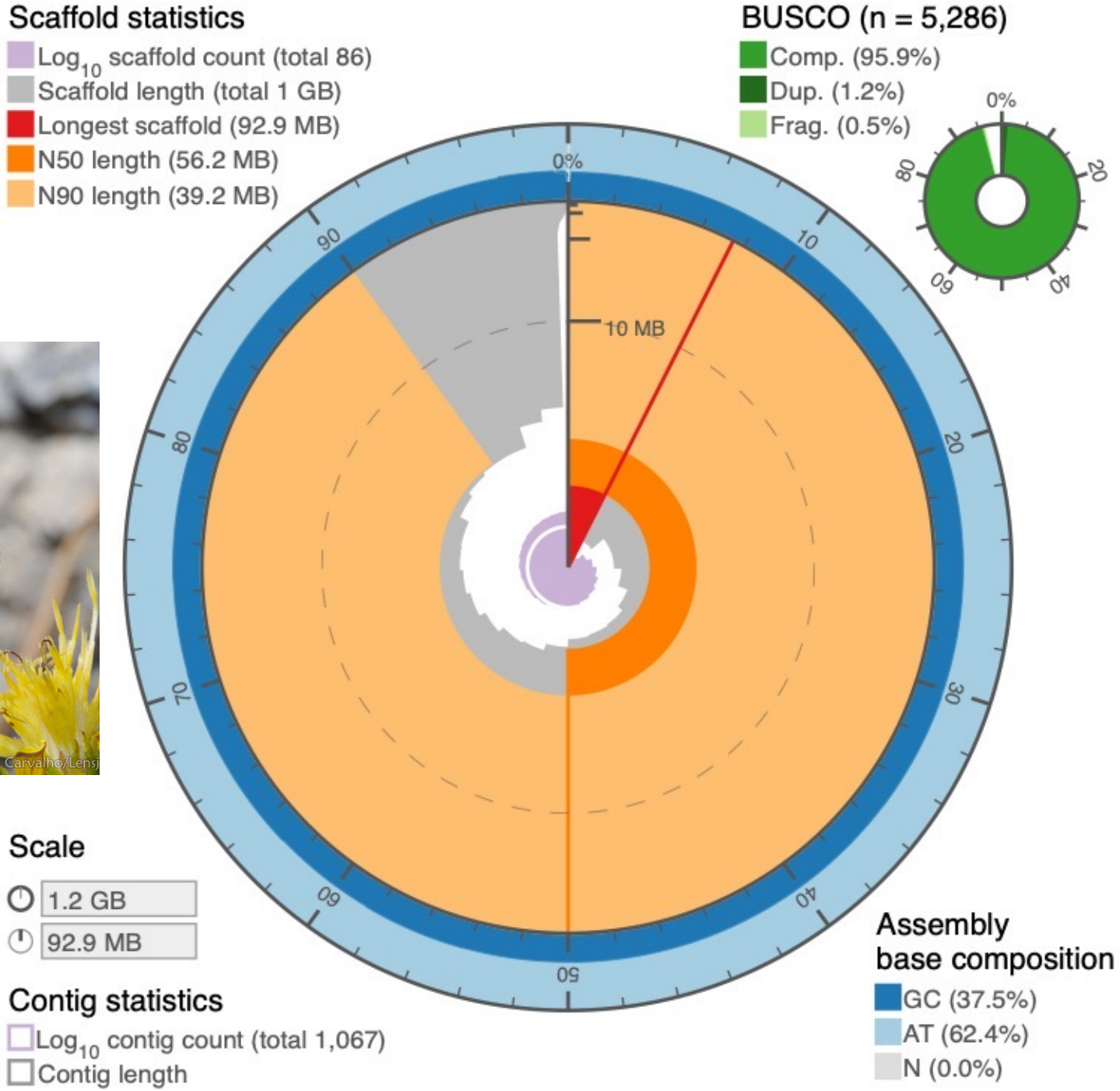


Assembly metrics

- No. of scaffolds/contigs
- Largest scaffold/contig
- N50 scaffold/contig size
  - 50% of genome contained in scaffolds/contigs of size ≥ N50
- L50
  - Minimum number of scaffolds/contigs with summed length ≥ 50% of genome
- Genome coverage (read coverage)
  - Each base represented by an average of X reads

“snail plot”


©Chris Carvalho/Lensj


Scaffold statistics
Log10 scaffold count (total 86)
Scaffold length (total 1 GB)
Longest scaffold (92.9 MB)
N50 length (56.2 MB)
N90 length (39.2 MB)
BUSCO (n = 5,286)
Comp. (95.9%)
Dup. (1.2%)
Frag. (0.5%)
0%
20
40
60
80
10 MB
Scale
1.2 GB
92.9 MB
Contig statistics
Log10 contig count (total 1,067)
Contig length
Assembly
base composition
GC (37.5%)
AT (62.4%)
N (0.0%)

fastq format

```
@M01478:6:000000000-A40C5:1:1101:16859:1439 1:N:0:1
ATCGTTTCGGAGCAAGGCAACTGTNTCAGGCACCATGAAGTTGAGCTATTCTACTGCGCCAACCTTTGCGAGA
GCCNTNNTTATCANCGTCAATTGGAANTCAGATGTGCCACCNNAAN
+
ABBBAABFBBBBGEGGFGGGGGHF#AAFF2AGFGHGHHHHHHFHFFDGFGHHHHGHEGGGGCGGGHFABEEGF
BFG#?##???FFH#??FEFGHHEHHG#??FFEDGGGFFHFH##??#
```

- Machine name
- Run number
- Flowcell ID
- Flowcell lane
- Tile in flowcell
- X-coordinate in tile
- Y-coordinate in tile
- Member of read pair (1/2)
- Read filtered? (Y/N)
- Control bits on (0 or even number)
- Sample number (when multiplexing)

```
[jdu282@mcc-login002 Bdor_pop_WGS_SRA_data]$ zcat SRR22045704_pass_1.fastq.gz | head
@SRR22045704.1.1 1 length=150
CNCCGGCTTTCGAAGCTTCTCCCTATTATAAAAAAAAAATTACAAACTAAAATATTTTTGTCCCCCTAGCTTGTATCAAACGATAT
ATTGTGCTACGGTTTGGCTTTGAATAAGCAAAATAATGTATTGGAACAGACGACGGAGATGGTT
+SRR22045704.1.1 1 length=150
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR22045704.2.1 2 length=150
TNGCGATCAGTTCCGCATGCCAGAACAAGCCAAGCAATTGGCCACATTCTTCTCGCTGTTCTATTTTGCTATTAATGCCGGTTCCA
CCATCTCGACCACCATTACACCCATACTCCGTGAGGATGTGCATTGCTTCGGCGATAAGAATTG
+SRR22045704.2.1 2 length=150
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFF
@SRR22045704.3.1 3 length=150
TCCATTATTTAGCAATCTTAGTTCTAAGAAAAACAAGTTTGTCTATAAATTATATTTCAAATTAAAATGAATATAAACAATCTTTC
ATTGCATATGGATTTTTTCTTATTTAAATTTAATAAACAACTAAGTAATATTATATAATAATAT
[jdu282@mcc-login002 Bdor_pop_WGS_SRA_data]$ zcat SRR22045704_pass_2.fastq.gz | head
@SRR22045704.1.2 1 length=150
TTAAGATTATTATTATAATTAGGATAAAAAACAGTGGTTAGTACGCAAAAACTTAGTTTGTATTTAAGTATATATTTGCATTTTGT
TAACATTAATACATTTTTATATAACAATGTTTTCTGTAATTTTAGGTTTATAATACTATGTAAG
+SRR22045704.1.2 1 length=150
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF:F
FFFFFFFFFFFFFFFFFFFFFF:FFFF:FFFFFFFFF:FF:FFFFFFFFFF::FFFFFF:
@SRR22045704.2.2 2 length=150
ACCACCAATAGCATCGGAAATACACTTTGACACACCGAATATCATGTTGCCAGCTGGTGGCTTAACCACGTAGAGCTTGCGGCCCG
CAACGAAAATGACCACCGAGAGCAACATAAGCATAGCCGGCACGCTGAAGGCGAGTGAAAAACA
+SRR22045704.2.2 2 length=150
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF
@SRR22045704.3.2 3 length=150
GTGGAATGATAAATTTAACTAACTTTACATTTCAAAGTACATAATTTCACTCAGGAGAACGGGTGCGGGGTCAGCTAGTATTTAAT
AAACACAAATAAGTAAAACTCTATGATAATATATCAAAATTTATATAATATTGGTGACTGAAGT
```

Phred quality score

A quality value Q is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

ASCII TABLE

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]
1	1	1	1	[START OF HEADING]
2	2	10	2	[START OF TEXT]
3	3	11	3	[END OF TEXT]
4	4	100	4	[END OF TRANSMISSION]
5	5	101	5	[ENQUIRY]
6	6	110	6	[ACKNOWLEDGE]
7	7	111	7	[BELL]
8	8	1000	10	[BACKSPACE]
9	9	1001	11	[HORIZONTAL TAB]
10	A	1010	12	[LINE FEED]
11	B	1011	13	[VERTICAL TAB]
12	C	1100	14	[FORM FEED]
13	D	1101	15	[CARRIAGE RETURN]
14	E	1110	16	[SHIFT OUT]
15	F	1111	17	[SHIFT IN]
16	10	10000	20	[DATA LINK ESCAPE]
17	11	10001	21	[DEVICE CONTROL 1]
18	12	10010	22	[DEVICE CONTROL 2]
19	13	10011	23	[DEVICE CONTROL 3]
20	14	10100	24	[DEVICE CONTROL 4]
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]
22	16	10110	26	[SYNCHRONOUS IDLE]
23	17	10111	27	[ENG OF TRANS. BLOCK]
24	18	11000	30	[CANCEL]
25	19	11001	31	[END OF MEDIUM]
26	1A	11010	32	[SUBSTITUTE]
27	1B	11011	33	[ESCAPE]
28	1C	11100	34	[FILE SEPARATOR]
29	1D	11101	35	[GROUP SEPARATOR]
30	1E	11110	36	[RECORD SEPARATOR]
31	1F	11111	37	[UNIT SEPARATOR]
32	20	100000	40	[SPACE]
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

Google “phred+33 quality score”

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

```
[jdu282@mcc-login002 Bdor_pop_WGS_SRA_data]$ zcat SRR22045704_pass_1.fastq.gz | head
@SRR22045704.1.1 1 length=150
CNCCGGCTTTCGAAGCTTCTCCCTATTATAAAAAAAAAATTACAAACTAAAATATTTTTGTCCCCCTAGCTTGTATCAAACGATAT
ATTGTGCTACGGTTTGGCTTTGAATAAGCAAAATAATGTATTGGAACAGACGACGGAGATGGTT
+SRR22045704.1.1 1 length=150
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR22045704.2.1 2 length=150
TNGCGATCAGTTCCGCATGCCAGAACAAGCCAAGCAATTGGCCACATTCTTCTCGCTGTTCTATTTTGCTATTAATGCCGGTTCCA
CCATCTCGACCACCATTACACCCATACTCCGTGAGGATGTGCATTGCTTCGGCGATAAGAATTG
+SRR22045704.2.1 2 length=150
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFF
@SRR22045704.3.1 3 length=150
TCCATTATTTAGCAATCTTAGTTCTAAGAAAAACAAGTTTGTCTATAAATTATATTTCAAATTAAAATGAATATAAACAATCTTTC
```

Quality trimming & masking

>Sequence1

16 16 21 9 10 13 14 12 8 8 9 16 24 21 19 19 19 25 25 33 35 35 34 34 34
34 34 34 34 40 45 45 56 56 56 51 51 40 45 37 37 37 40 40 40 40 40 40 39 39
39 40 40 40 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 51 45 39 39
39 39 39 39 39 39 39 39 39 39 39 40 51 51 51 51 51 56 56 56 56 56 56 56
40 35 35 35 35 35 39 40 40 45 45 45 45 45 51 56 40 39 39 39 39 39 45 45 45
45 40 40 40 39 35 35 40 40 40 40 40 40 38 38 38 39 25 23 18 10 8 9 23 31
51 51 45 43 43 43 43 43 43 43 43 43 43 43 43 56 56 56 56 56 56 56 56 43 43
43 43 43 43 45 45 45 51 51 56 51 51 51 51 51 51 56 51 45 43 43 56 56 56 56
56 56 51 51 51 45 45 45 40 40 40 44 42 38 38 40 40 40 51 56 56 56 56 56 46
46 51 51 51 51 56 56 56 56 51 40 45 45 40 40 40 42 42 42 45 45 42 42 42 42
56 42 42 40 40 34 37 33 40 40 40 44 48 48 48 29 29 29 26 32 29 32 32 32 32
33 44 48 56 40 40 40 40 40 40 40 40 40 37 34 34 37 40 40 40 40 37 34 34 48
40 32 28 25 25 25 34 48 48 48 40 40 32 29 24 25 29 40 40 40 40 40 40 33 33
37 40 40 40 43 43 42 42 42 44 44 56 56 56 56 56 40 35 34 33 33 40 40 40 40
40 40 29 29 34 29 29 29 29 40 29 34 25 27 23 23 21 23 18 20 25 25 25 32 32
32 32 29 18 20 14 16 16 17 17 16 22 20 18 25 19 14 16 15 26 27

Trimmed & masked sequence

>Sequence1

CCAGAAACTACGCGGTGGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGATCGTC

CGCCAGACTAAAGAAGTCCAAGAGTTGGCTCGCCAAAACGCGCTAAAAACGCAAAAAGCGG

CGACCAGTAGA**NNNN**AGGCGAGGCAGGAAGAACAAGCCAACTTTTGGGGTTAACGACTATG

TTTTCGTCAAGAAAAAAGGGTTTCCGACGACCGCACCGACGACCAGATTGGATTCACAGTG

GACCGGACCATGGCAGATTCTAGAAGAACGAGGATATAGCTATGTTTTGGACGTACCTGAA

TCGTTTAAAGGAAAAAATTTGTTCCACGCAGACCGCCTCCGCAAAGCCGCAATGGACCCAT

TACCACAACAGAAAAGAGAGCCGCCTCCGCCAGAAGAGATCAACGCCAGAGTTTGTGGTCG

ATAAAGTTTTAGCGTCCCGATTATTTGGCCGGAGTAAGATATTGCAATACCAGGTCGCATG

GCAAGGATGTGATCCAGACGACACGTGGTACCCGGCTGAAAACTTCAAGAATTCAGCGACA

GCCCTTGACGACTTCCACAAGAAGTAC

fasta format

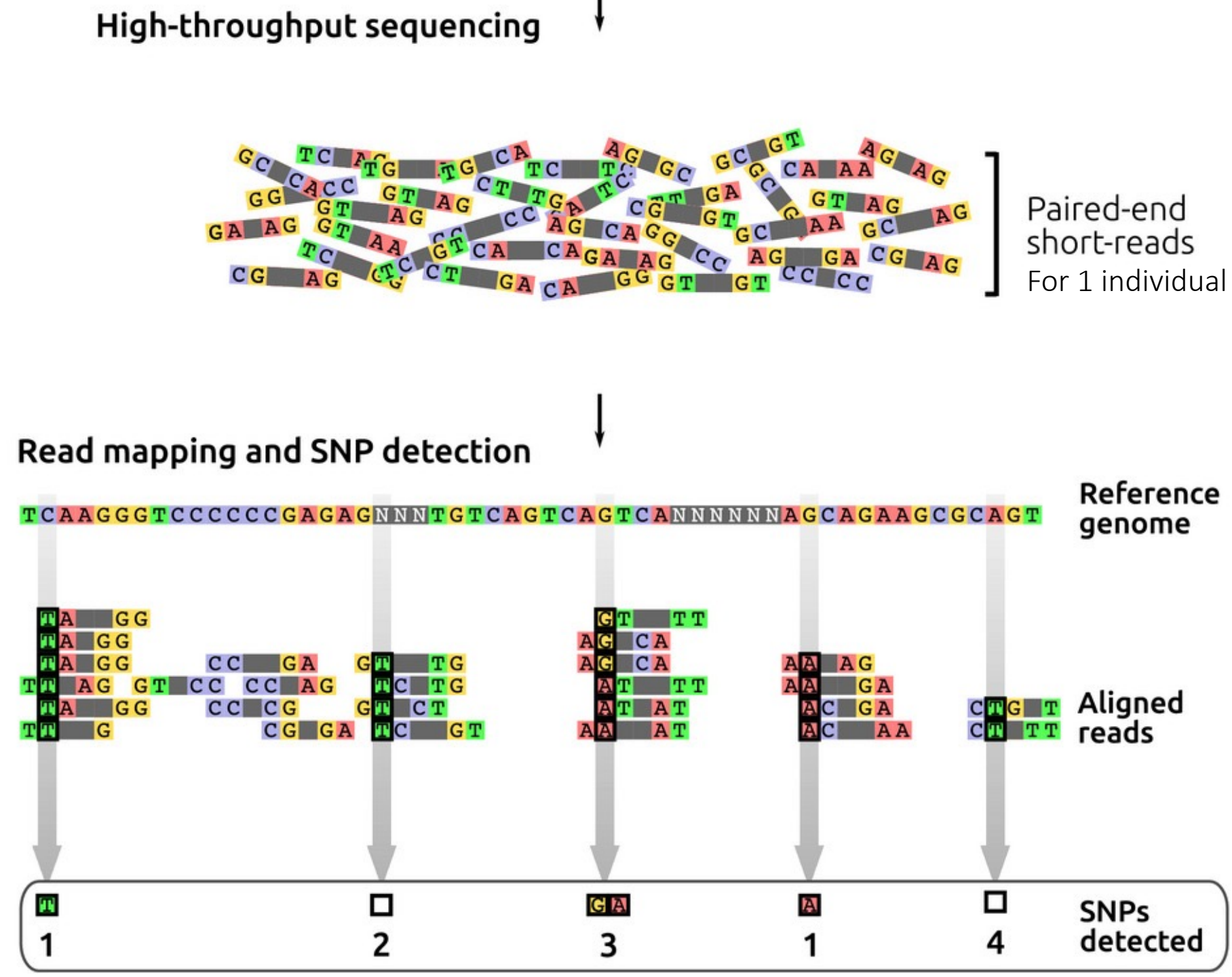
```
>header
[sequence]

>geneXYZ
ATTTGCGATAGCTAGCTCGATCGATCG

>geneXYZ
ATTTGCGATAGCTAGC
TCGATCGATCG
```

fasta format

```
[jdu282@login003 6_fastas]$ head OG0008033_73-548_s163_1.renamed.fasta
>fraterculus_Mexico_KST_002 OG0008033_73-548_s163_1 matched_4 trashed_0
ATGTCCGATCTTTACGATATGCAAGTACAAAACCATGAACAAAATGAAGTATTCAAGTGTTTTAGCGCATGCGGTGGCTTTGGCGGACTTCAAACTGTAAAT
ACCACTTCAGAATCTCGGGAGCAAATCGAAGAAGAATATGAAAAGTACAAAAATACTCGCATTGATTTCAATCGCTTACGTACATTGAAAGATCTTGGTATC
GATATAGACTTTCTTGATAATATCGAGGCAGAAATGAAAAATTTCGAATTGACTCGGCGTCTTCAGGAGCATTTAAGCAATAATCTAAATTTGATAGAGAAA
CTGCGCGCCACGCAACATGAACGACTCTCTCAGCCACTGCCGCAGCACCTGGCCTATGTACCGCAAGCTAGTGCCGAAGAAGTGCAT
>fraterculus_Mexico_KST_003 OG0008033_73-548_s163_1 matched_1 trashed_1
ATGTCCGATCTTTACGATATGCAAGTACAAAACCATGAACAAAATGAAGTATTCAAGTGTTTTAGCGCATGCGGTGGCTTTGGCGGACTTCAAACTGTAAAT
ACCACTTCAGAATCTCGGGAGCAAATCGAAGAAGAATATGAAAAGTACAAAAATACTCGCATTGATTTCAATCGCTTACGTACATTGAAAGATCTTGGTATC
GATATAGACTTTCTTGATAATATCGAGGCAGAAATGAAAAATTTCGAATTGACTCGGCGTCTTCAGGAGCATTTAAGCAATAATCTAAATTTGATAGAGAAA
CTGCGCGCCACGCAACATGAACGACTCTCTCAGCCACTGCCGCAGCACCTGGCCTATGTACCGCAAGCTAGTGCCGAAGAGTGCAT
>fraterculus_Brazil_KST_004 OG0008033_73-548_s163_1 matched_1 trashed_0
ATGTCCGATCTTTACGATATGCAAGTACAAAACCATGAACAAAATGAAGTATTCAAGTGTTTTAGCGCATGCGGTGGCTTTGGTGGACTTCAAACTGTAAAT
ACCACTTCAGAATCTCGGGAGCAAATCGAAGAAGAATATGAAAAGTACAAAAATACTCGCATTGATTTCAATCGCTTACGTACATTGAAAGATCTTGGTATC
GATATAGACTTTCTTGATAATATCGAGGCAGAAATGAAAAATTTCGAATTGACTCGGCGTCTTCAGGAGCATTTAAGTAATAATCTAAATTTGATAGAGAAA
CTGCGCGCCACGCAACATGAACGACTCTCTCAGCCACTGCCGCAGCACCTGGCCTATGTACCGCAAGCTAGTGCCGAAGAAGTGCAT
>fraterculus_Brazil_Vacaria_KST_006 OG0008033_73-548_s163_1 matched_1 trashed_0
ATGTCCGATCTTTACGATATGCAAGTACAAAACCATGAACAAAATGAAGTGTTCAAGTGTTTTAGCGCATGCGGTGGCTTTGGTGGACTTCAAACTGTAATA
AATACCACTTCAGAATCTCGGGAGCAAATCGAAGAAGAATATGAAAAGTACAAAAATACTCGCATTGATTTCAATCGCTTACGTACATTGAAAGATCTTGGT
ATCGATATAGACTTTCTTGATAATATCGAGGCAGAAATGAAAAATTTCGAATTGACTCGGCGTCTTCAGGAGCATTTAAGCAATAATCTAAATTTGATAGAG
AAACTGCGCGCTACGCAGCATGAGAGACTCTCTCAGCCACTGCCGCAGCACCTGGCCTATGTACCGCAAGCTAGTGCCGAAGAAGTGCAT
>fraterculus_Peru_Libertad_KST_016 OG0008033_73-548_s163_1 matched_1 trashed_0
ATGTCCGATCTTTACGATATGCAAGTACAAAACCATGAACAAAATGAAGTATTCAAGTGTTTTAGCGCATGCGGTGGCTTTGGTGGACTTCAAACTGTTAAT
ACCACTTCAGAATCTCGGGAGCAAATCGAAGAAGAATATGAAAAGTACAAAAATACTCGCATTGATTTCAATCGCTTACGTACATTGAAAGATCTTGGTATC
GATATAGACTTTCTTGATAATATCGAGGCAGAAATGAAAAATTTCGAATTGACTCGGCGTCTTCAGGAGCATTTAAGCAATAATCTAAATTTGATAGAGAAA
CTGCGCGCCACGCAACATGAGAGACTCTCTCAGCCACTGCCGCAGCACCTGGCCTATGTACCGCAAGCTAGTGCCGAAGAAGTGCAT
```


High-throughput sequencing
Paired-end
short-reads
For 1 individual
Read mapping and SNP detection
Reference
genome
Aligned
reads
SNPs
detected
1
2
3
1
4

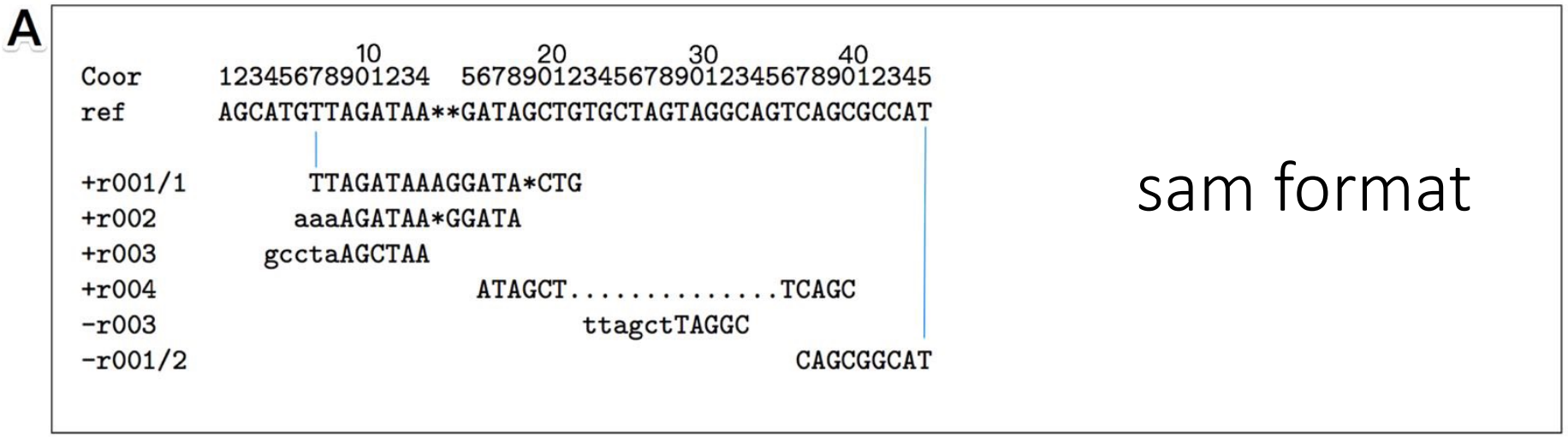
julian — jdu282@mcc-login002:/scratch/jdu282/dorsalis_wgs/2_bwa2_out — ssh jdu282@mcc.uky.edu — 206×64

GNU nano 2.9.8 SRR22045704.paired.sam

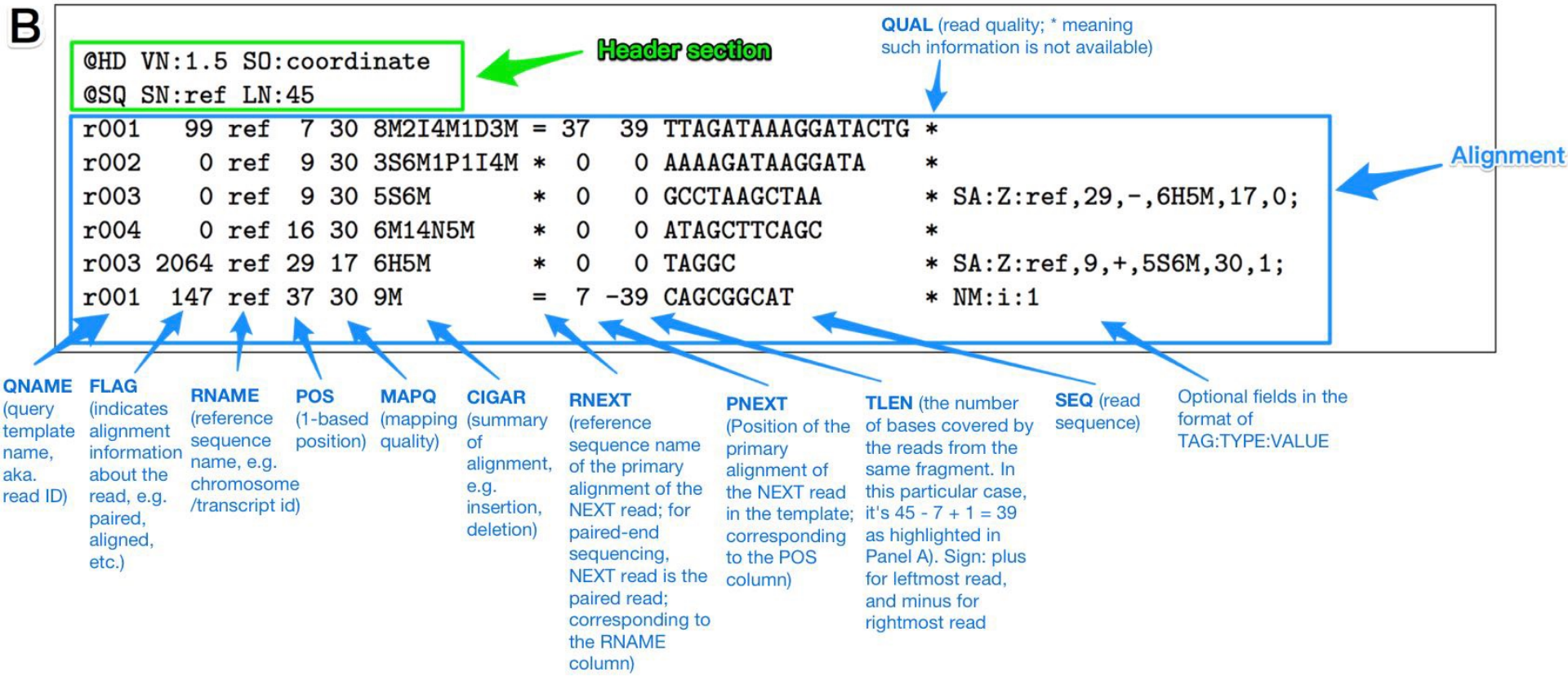
sam format

```
@SQ     SN:Chromosome1  LN:114943187
@SQ     SN:Chromosome2  LN:106640461
@SQ     SN:Chromosome3  LN:93297747
@SQ     SN:Chromosome4  LN:74755765
@SQ     SN:Chromosome5  LN:74733959
@SQ     SN:Chromosome6  LN:42482483
@PG     ID:bwa-mem2     PN:bwa-mem2     VN:2.2.1        CL:bwa-mem2 mem -t 20 -M ./genome.fasta ./1_fastp_out/SRR22045704.R1.fastpOut.fq.gz ./1_fastp_out/SRR22045704.R2.fastpOut.fq
SRR22045704.1.1 65      Chromosome1     66796837        60      150M    =       66796837        0       CNCCGGCTTTCGAAGCTTCTCCCTATTATAAAAAAAAAATTACAAACTAAAATATTTTTGTCCCCCTAGCTTGTAT
SRR22045704.1.1 129     Chromosome1     66796837        60      150M    =       66796837        0       CNCCGGCTTTCGAAGCTTCTCCCTATTATAAAAAAAAAATTACAAACTAAAATATTTTTGTCCCCCTAGCTTGTAT
SRR22045704.2.1 113     Chromosome4     71064415        60      150M    =       71064415        0       CAATTCTTATCGCCGAAGCAATGCACATCCTCACGGAGTATGGGTGTAATGGTGGTCGAGATGGTGGAACCGGCAT
SRR22045704.2.1 177     Chromosome4     71064415        60      150M    =       71064415        0       CAATTCTTATCGCCGAAGCAATGCACATCCTCACGGAGTATGGGTGTAATGGTGGTCGAGATGGTGGAACCGGCAT
SRR22045704.3.1 65      Chromosome3     39108034        60      150M    =       39108034        0       TCCATTATTTAGCAATCTTAGTTCTAAGAAAAACAAGTTTGTCTATAAATTATATTTCAAATTAAAATGAATATAA
SRR22045704.3.1 129     Chromosome3     39108034        60      150M    =       39108034        0       TCCATTATTTAGCAATCTTAGTTCTAAGAAAAACAAGTTTGTCTATAAATTATATTTCAAATTAAAATGAATATAA
SRR22045704.4.1 97      Chromosome3     46932321        0       150M    =       46806818        -125355 GNCGATTTTTCGAAATTTCTTCCATTAGCTTATAAATAATGATGCGTAGTATGCTCCTGGTATAGTTTTTCCTTT
SRR22045704.4.1 145     Chromosome3     46806818        0       150M    =       46932321        125355  GTGTATTCGGCTGGGAAAGTGATGGCGAGTGTTTTTTTGGGACAGCCATGGAATTATTTTTATCGACTATCTTGAAA
SRR22045704.5.1 65      Chromosome2     19445008        60      150M    =       19445008        0       ANTTCATTGAGACATACCACTCCCCGTTCGATTAGTTCTTCAACGGCGCGAGCCACACCTGCCTCATCGCCAGTCC
SRR22045704.5.1 129     Chromosome2     19445008        60      150M    =       19445008        0       ANTTCATTGAGACATACCACTCCCCGTTCGATTAGTTCTTCAACGGCGCGAGCCACACCTGCCTCATCGCCAGTCC
SRR22045704.6.1 77      *       0       0       *       *       0       0       CNCAAATCTTATCTCTTATGTTAAGATGAGCCTTTATGTGATCTTGCAAATCGCGTGCTTTCAAAAAATCTTTCGGGCACATAGTACATTTGAATGCTCT
SRR22045704.6.1 141     *       0       0       *       *       0       0       CNCAAATCTTATCTCTTATGTTAAGATGAGCCTTTATGTGATCTTGCAAATCGCGTGCTTTCAAAAAATCTTTCGGGCACATAGTACATTTGAATGCTCT
SRR22045704.7.1 65      Chromosome2     2337793 60      96M1D54M        =       2337793 0       GNTAACGTTAGTTTTCTTTTTCAGTAATTTTTCAAATGTTGTCGATAAGGTGCATTGGAAAACGACAGAATAATATTCGTAGAT
SRR22045704.7.1 129     Chromosome2     2337793 60      96M1D54M        =       2337793 0       GNTAACGTTAGTTTTCTTTTTCAGTAATTTTTCAAATGTTGTCGATAAGGTGCATTGGAAAACGACAGAATAATATTCGTAGAT
SRR22045704.8.1 65      Chromosome1     2913859 60      150M    =       2913859 0       CNACTAACCTAATATTTCCCAAAACTAACCATAAGTGGACTCGTTGGATTATTTATACTATTTTTCCCTACATTTACCACTGGTACCTAATA
SRR22045704.8.1 129     Chromosome1     2913859 60      150M    =       2913859 0       CNACTAACCTAATATTTCCCAAAACTAACCATAAGTGGACTCGTTGGATTATTTATACTATTTTTCCCTACATTTACCACTGGTACCTAATA
SRR22045704.9.1 113     Chromosome6     40399260        4       150M    =       40399260        0       CGAGTGACCCTAGCGCAACTTCGTTCTGGGTACTGTAGCAGGTTAAACTCCTACTTATCCAGAATCGACCCCGACA
SRR22045704.9.1 177     Chromosome6     40399260        4       150M    =       40399260        0       CGAGTGACCCTAGCGCAACTTCGTTCTGGGTACTGTAGCAGGTTAAACTCCTACTTATCCAGAATCGACCCCGACA
SRR22045704.10.1        77      *       0       0       *       *       0       0       TNTGTTTGCTTGCGGTTTTTCAGCTGGTACAAGCGCTATTCGCCACCTTCCCCCGCCTTAAACCTTTCGGGCAGACTTTGGTTTTTATGCGA
SRR22045704.10.1        141     *       0       0       *       *       0       0       TNTGTTTGCTTGCGGTTTTTCAGCTGGTACAAGCGCTATTCGCCACCTTCCCCCGCCTTAAACCTTTCGGGCAGACTTTGGTTTTTATGCGA
SRR22045704.11.1        65      Chromosome5     4910154 60      150M    =       4910154 0       CNCTCTTCGGATATAATAAAATCGATCACCACTAACTTCTGGAGTCATCTATAGACAAACAGCGGTAACTTAGTAGCGCACCTA
SRR22045704.11.1        129     Chromosome5     4910154 60      150M    =       4910154 0       CNCTCTTCGGATATAATAAAATCGATCACCACTAACTTCTGGAGTCATCTATAGACAAACAGCGGTAACTTAGTAGCGCACCTA
SRR22045704.12.1        65      Chromosome4     35578243        3       2S73M3D34M1I40M =       35578243        0       CNTTTTTTAAGATTTTCATCAGTTGAAGAGGTCGAAGGACGTCCAGAACGAGACATGTCT
SRR22045704.12.1        129     Chromosome4     35578243        3       2S73M3D34M1I40M =       35578243        0       CNTTTTTTAAGATTTTCATCAGTTGAAGAGGTCGAAGGACGTCCAGAACGAGACATGTCT
SRR22045704.13.1        65      Chromosome5     34861844        60      9S141M  =       34861844        0       CNGGAACTGTAGTAACATATGAGCTGATGCTTATATAATTCAATCAGGATAGCAAACTTGCAGCATAG
SRR22045704.13.1        129     Chromosome5     34861844        60      9S141M  =       34861844        0       CNGGAACTGTAGTAACATATGAGCTGATGCTTATATAATTCAATCAGGATAGCAAACTTGCAGCATAG
SRR22045704.14.1        113     Chromosome1     100079789       0       150M    =       100079789       0       ATATGCCCCTTCTCTCACATGGAACTGAACTCTTGTTCTTACCCATTTTTTTTGAAAACTTTCAAAAA
SRR22045704.14.1        177     Chromosome1     100079789       0       150M    =       100079789       0       ATATGCCCCTTCTCTCACATGGAACTGAACTCTTGTTCTTACCCATTTTTTTTGAAAACTTTCAAAAA
SRR22045704.15.1        113     Chromosome6     39741252        0       150M    =       40571965        830714  TTTCGTGCATTTATTGTTAGACATATAAACCGAAAAGGGATTGTAGTGTCCGCAAGGTCCATGAACCA
SRR22045704.15.1        177     Chromosome6     40571965        0       150M    =       39741252        -830714 TTTCGTGCATTTATTGTTAGACATATAAACCGAAAAGGGATTGTAGTGTCCGCAAGGTCCATGAACCA
SRR22045704.16.1        65      Chromosome6     5139136 0       150M    =       4241163 -897974 CCATCTTGTTGAAACCACATGTCAACCAAGTTCAGTTCTTCCATTTTTGGCAACAAAAAGTTTGTTAGCATCGAACGATAGCGA
SRR22045704.16.1        129     Chromosome6     4241163 0       150M    =       5139136 897974  CCATCTTGTTGAAACCACATGTCAACCAAGTTCAGTTCTTCCATTTTTGGCAACAAAAAGTTTGTTAGCATCGAACGATAGCGA
SRR22045704.17.1        65      Chromosome5     15646746        60      150M    =       15646746        0       GACCTCCCTTCTCCTCATCTTTTCCCCTTTCTCTCAGCTTTTGCCAGTGGAAATATTTTCTTTTAAAG
SRR22045704.17.1        129     Chromosome5     15646746        60      150M    =       15646746        0       GACCTCCCTTCTCCTCATCTTTTCCCCTTTCTCTCAGCTTTTGCCAGTGGAAATATTTTCTTTTAAAG
SRR22045704.18.1        65      Chromosome1     5316626 60      95M3I4M2I30M16S =       5316626 0       GCGGTATCTGGGGGCAAGAGAAGTCATTGCTGCGACAATTTTTAATCGTATGCTTAATATAGCAAATTCTTTTTTA
SRR22045704.18.1        129     Chromosome1     5316626 60      95M3I4M2I30M16S =       5316626 0       GCGGTATCTGGGGGCAAGAGAAGTCATTGCTGCGACAATTTTTAATCGTATGCTTAATATAGCAAATTCTTTTTTA
SRR22045704.19.1        113     Chromosome2     85734025        60      150M    =       85734025        0       AGCGAGCAGCACCAAAATTGCACAGACGTCGACGAAAACCATGGGCAAGCACCCATCAGAATCAAAGC
SRR22045704.19.1        177     Chromosome2     85734025        60      150M    =       85734025        0       AGCGAGCAGCACCAAAATTGCACAGACGTCGACGAAAACCATGGGCAAGCACCCATCAGAATCAAAGC
SRR22045704.20.1        113     Chromosome2     6653562 60      150M    =       6653562 0       TCTGTGGACTTCTCTCCGGATGCCGCACACCGCAGTAAGTTCCATATTGTAATTGCAAATGAAGTGTTAAATTGTTTTTGATGA
SRR22045704.20.1        177     Chromosome2     6653562 60      150M    =       6653562 0       TCTGTGGACTTCTCTCCGGATGCCGCACACCGCAGTAAGTTCCATATTGTAATTGCAAATGAAGTGTTAAATTGTTTTTGATGA
SRR22045704.21.1        65      Chromosome5     50095269        60      112M4D38M       =       50095269        0       AAATGTACTAGGCTAGAAAGGAACAACAAAAGATAATTTATATCAAGACAAATACAATTT
SRR22045704.21.1        129     Chromosome5     50095269        60      112M4D38M       =       50095269        0       AAATGTACTAGGCTAGAAAGGAACAACAAAAGATAATTTATATCAAGACAAATACAATTT
SRR22045704.22.1        65      Chromosome1     75949137        60      150M    =       75949137        0       CAGGGCTCTGACAGTGATGGTACTCTCGTTGTGACATATGGTCACCATGGCTGTACGTAAGAGCATAT
SRR22045704.22.1        129     Chromosome1     75949137        60      150M    =       75949137        0       CAGGGCTCTGACAGTGATGGTACTCTCGTTGTGACATATGGTCACCATGGCTGTACGTAAGAGCATAT
SRR22045704.23.1        65      Chromosome3     67673592        60      7S90M1I52M      =       67673592        0       ANATTCCTCTAAAGACTATCATCACAATAGTTTTATTTTCATTTGAATCTTTCTCGGGTA
SRR22045704.23.1        129     Chromosome3     67673592        60      7S90M1I52M      =       67673592        0       ANATTCCTCTAAAGACTATCATCACAATAGTTTTATTTTCATTTGAATCTTTCTCGGGTA
SRR22045704.24.1        65      Chromosome2     44807853        60      150M    =       44807853        0       TNTAAATTTTATTATAACTTGATTGAATTAATAAGCAGAATAATCCATTTATGTAAATGCTTGTAAAA
SRR22045704.24.1        129     Chromosome2     44807853        60      150M    =       44807853        0       TNTAAATTTTATTATAACTTGATTGAATTAATAAGCAGAATAATCCATTTATGTAAATGCTTGTAAAA
SRR22045704.25.1        65      Chromosome5     52366156        60      150M    =       52366156        0       ANTTCAATAAAAATAAAAAATTCAATAAAAATACCGCAAGACTTTTTTGACAACCTAATATGTTAGAA
SRR22045704.25.1        129     Chromosome5     52366156        60      150M    =       52366156        0       ANTTCAATAAAAATAAAAAATTCAATAAAAATACCGCAAGACTTTTTTGACAACCTAATATGTTAGAA
SRR22045704.26.1        113     Chromosome3     38956422        0       150M    =       39433155        476734  ATGAGCCTCATCGCTGAACAAAATTAAAGCGCGAAACACATTTCGAACCGAACACTGATTTTGGTAAT
SRR22045704.26.1        177     Chromosome3     39433155        0       150M    =       38956422        -476734 ATGAGCCTCATCGCTGAACAAAATTAAAGCGCGAAACACATTTCGAACCGAACACTGATTTTGGTAAT
```

^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify ^C Cur Pos M-U Undo M-A Mark Text M-] To Bracket M-A Previous ^B Back
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Spell ^_ Go To Line M-E Redo M-6 Copy Text M-W WhereIs Next M-▼ Next ^F Forward


A
10 20 30 40
Coor 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1 TTAGATAAAGGATA*CTG
+r002 aaaAGATAA*GGATA
+r003 gcctaAGCTAA
+r004 ATAGCT..............TCAGC
-r003 ttagctTAGGC
-r001/2 CAGCGGCAT

sam format


B
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
Header section
QUAL (read quality; * meaning such information is not available)
r001 99 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 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
Alignment
QNAME (query template name, aka. read ID)
FLAG (indicates alignment information about the read, e.g. paired, aligned, etc.)
RNAME (reference sequence name, e.g. chromosome /transcript id)
POS (1-based position)
MAPQ (mapping quality)
CIGAR (summary of alignment, e.g. insertion, deletion)
RNEXT (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)
PNEXT (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)
TLEN (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read
SEQ (read sequence)
Optional fields in the format of TAG:TYPE:VALUE

CIGAR string

```
Ref   CTGGCCATTATCTC--GGTGGTAGGACATGGCATGCCC
Read        aaATGTCGCGGTG.TAGGAggatcc
```

2S|5M|2I|4M|1D|5M|6S

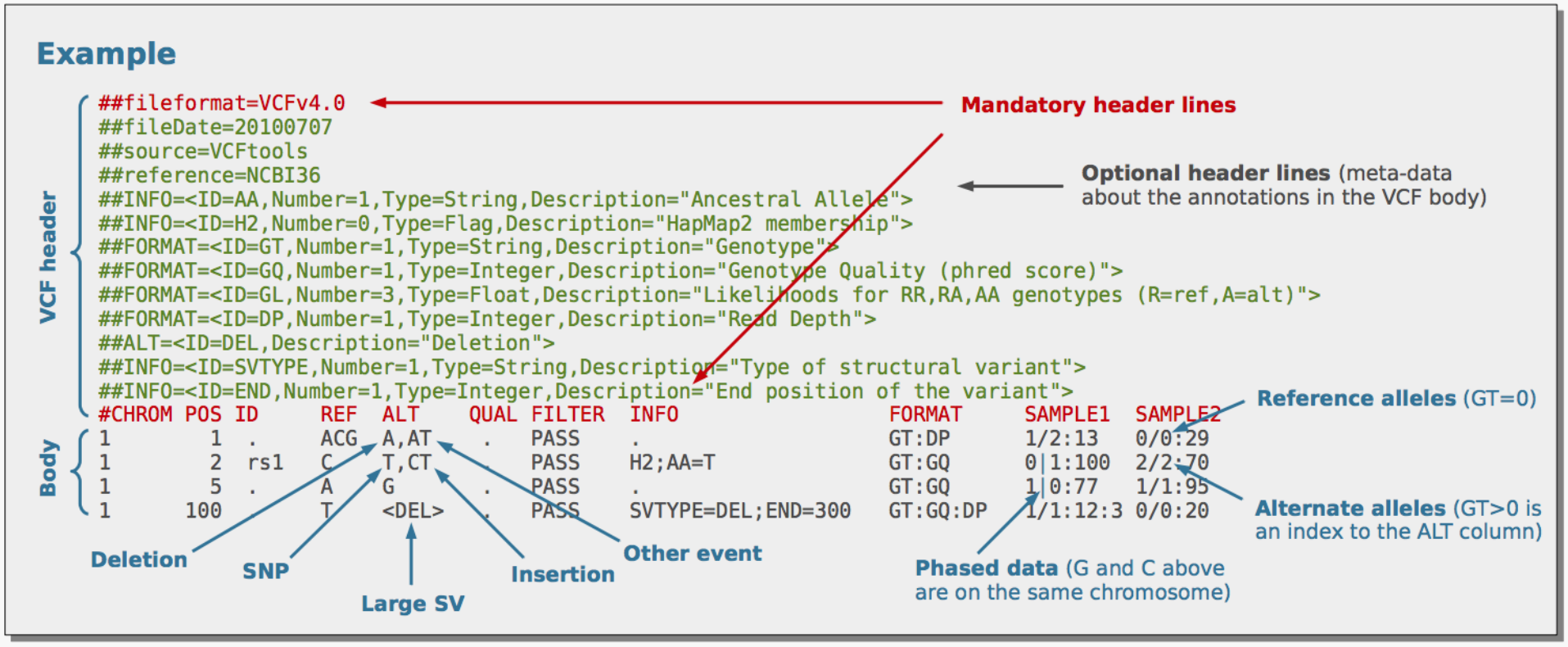
	Op	BAM	Description
	M	0	alignment match (can be a sequence match or mismatch)
	I	1	insertion to the reference
	D	2	deletion from the reference
	N	3	skipped region from the reference
	S	4	soft clipping (clipped sequences present in SEQ)
*	H	5	hard clipping (clipped sequences NOT present in SEQ)
*	P	6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

"N" indicates splicing event in RNAseq BAMs

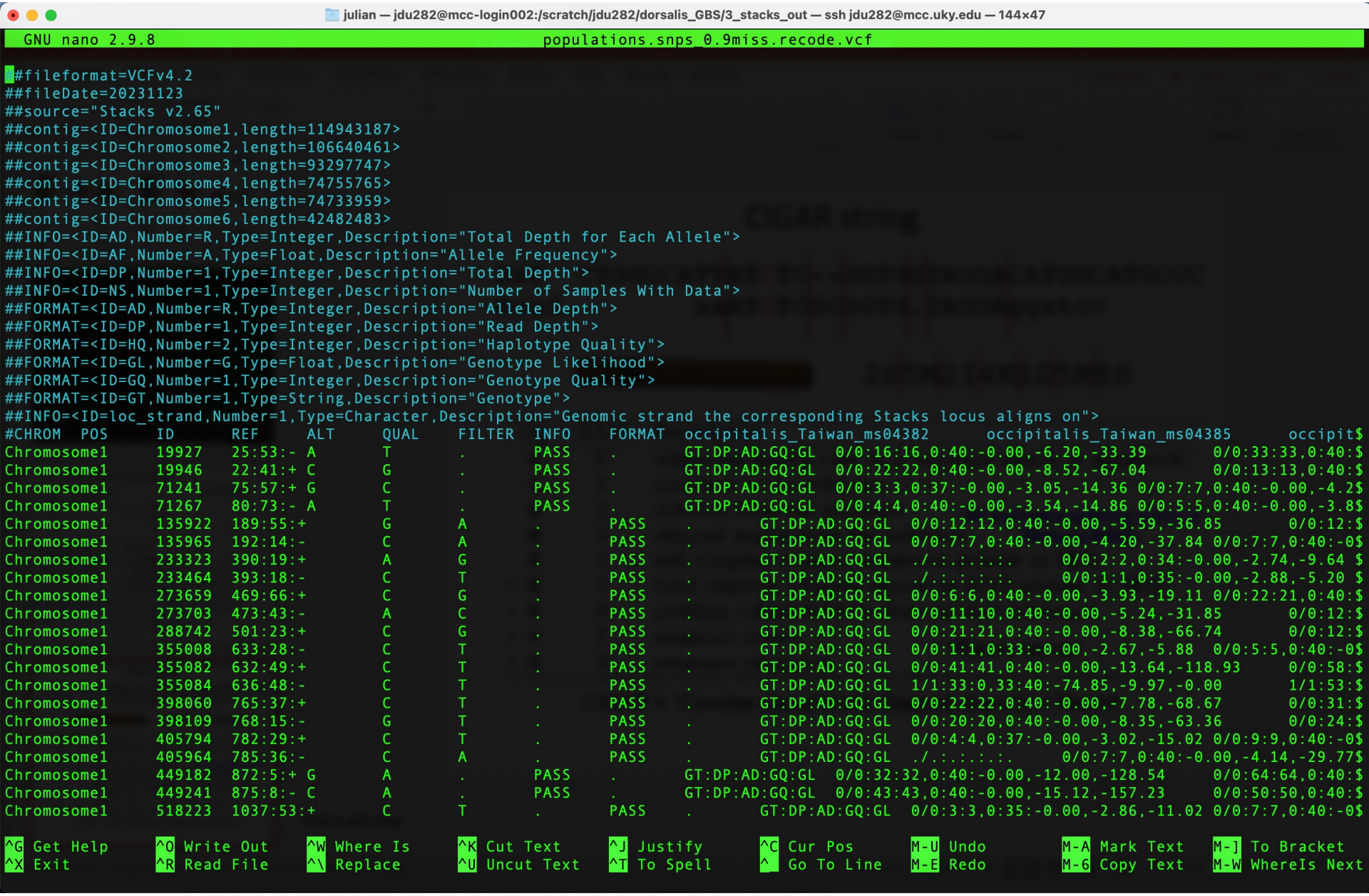
*Rarer / newer

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

vcf format


Example
VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##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 (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##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 . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

vcf format via Stacks (GBS)



vcf format via GATK (WGS)