

1/28

Pair 1 = Read 1 + Read 2

Pair 2 = Read 3 + Read 4

Pair 3 = Read 5 + Read 6

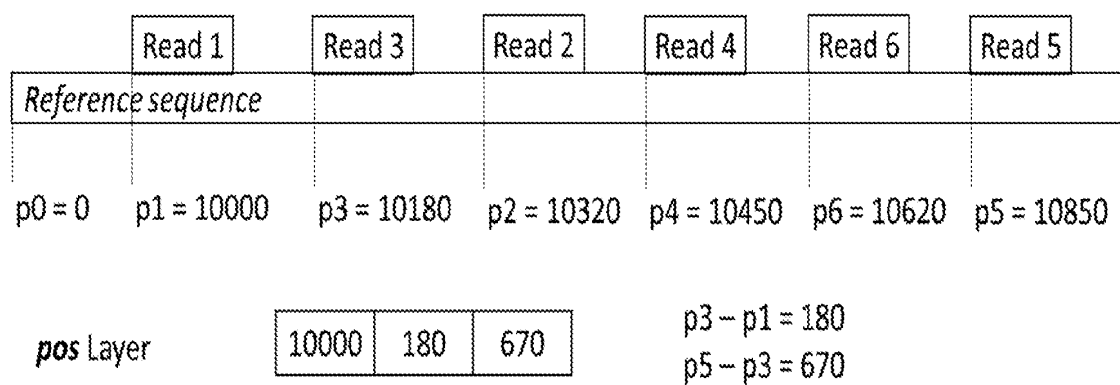
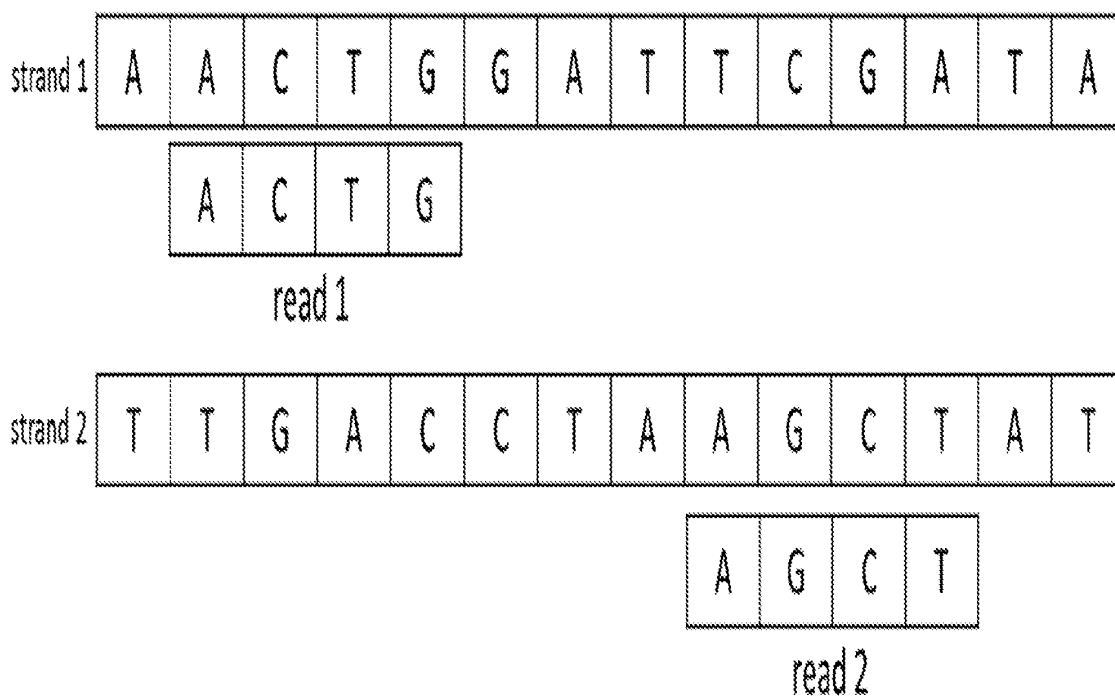


Figure 1



5

Figure 2

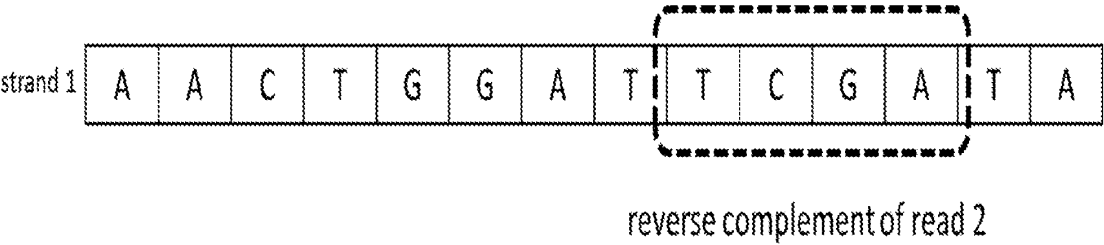


Figure 3

5

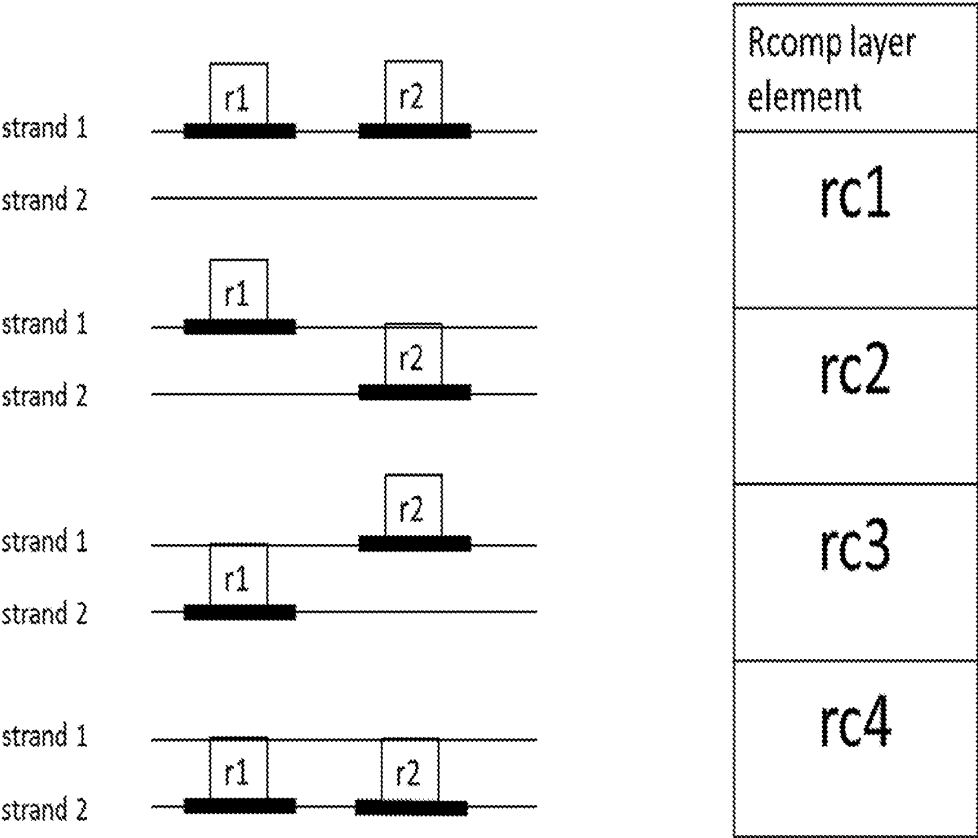


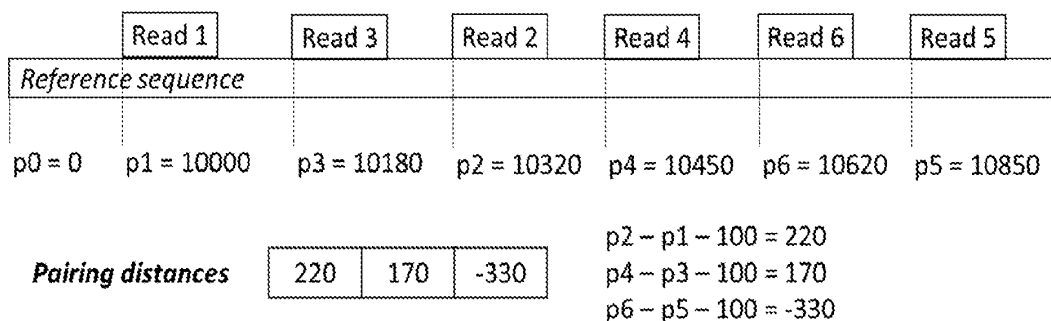
Figure 4

10

3/28

EXAMPLE

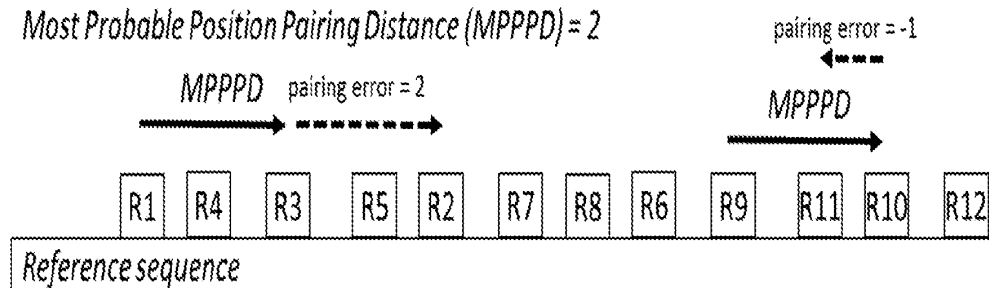
Constant reads length = 100

**Figure 5**

5

EXAMPLE

Most Probable Position Pairing Distance (MPPPD) = 2



pair

2	-3	2	-1	0	0				
---	----	---	----	---	---	--	--	--	--

Pairing errors vector

Optionally the pair layer can be differentially encoded into pair'

pair'

2	-5	5	-3	1	0				
---	----	---	----	---	---	--	--	--	--

Pairing errors vector

Figure 6

4/28

Pair 1 = Read 1 + Read 2

Pair 2 = Read 3 + Read 4

Pair 3 = Read 5 + Read 6

Read 1

Read 3

Read 2

Read 6

Read 5

Reference sequence 1

Read 4

Reference sequence 4

p0=0 p1=10000 p3=10180 p2=10320 p4=10450 p6=10620 p5=10850

Pairing distances

220

0xffffffff

4

170

-330

reserved value (flag)

reference sequence id if different from current reference

$$p_2 - p_1 - 100 = 220$$
$$p_4 - p_3 - 100 = 170$$
$$p_6 - p_5 - 100 = -330$$

Figure 7

5

Ref. A A C T G G A T T C G A T A

Read 1 A N T N

Read 2 N G A N

nmis layer 1 3 4 7 S

nmis layer (differential enc.) 1 2 1 3 S

Figure 8

10

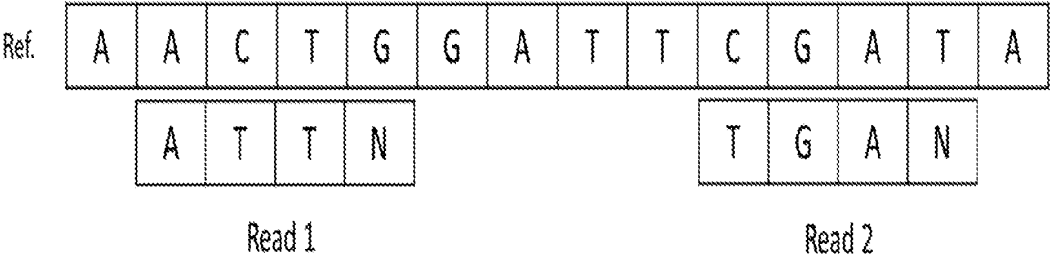


Figure 9

5

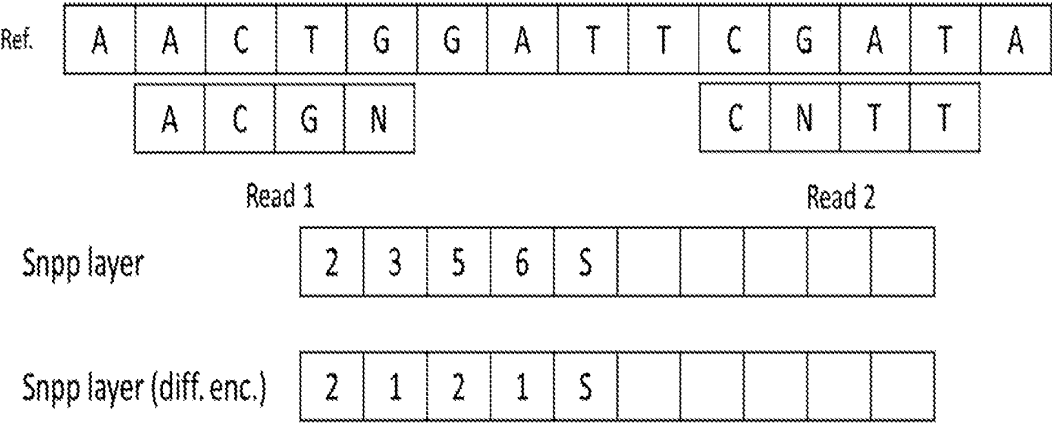


Figure 10

10

15

6/28

snpt Layer (without IUPAC codes)

A substitution type is calculated as index of a substitutions vector composed by all the possible symbols. For example:

$S = [A, C, G, T, N, Z]$ **where Z = deletion**

Index direction

- ENCODING is from right to left
- DECODING from left to right

Ref	Read	Encoded Symb.
A	del.	$\text{idx}(A,Z) = 5$
C	del.	$\text{idx}(C,Z) = 4$
G	del.	$\text{idx}(G,Z) = 3$
T	del.	$\text{idx}(T,Z) = 2$

Ref	Read	Encoded Symb.
N	A	$\text{idx}(N,A) = 2$
N	C	$\text{idx}(N,C) = 3$
N	G	$\text{idx}(N,G) = 4$
N	T	$\text{idx}(N,T) = 5$

Ref	Read	Encoded Symb.
A	C	$\text{idx}(A,C) = 1$
A	G	$\text{idx}(A,G) = 2$
A	T	$\text{idx}(A,T) = 3$
A	N	$\text{idx}(A,N) = 4$
C	A	$\text{idx}(C,A) = 5$
C	G	$\text{idx}(C,G) = 1$
C	T	$\text{idx}(C,T) = 2$
C	N	$\text{idx}(C,N) = 3$
G	A	$\text{idx}(G,A) = 4$
G	C	$\text{idx}(G,C) = 5$
G	T	$\text{idx}(G,T) = 1$
G	N	$\text{idx}(G,N) = 2$
T	A	$\text{idx}(T,A) = 3$
T	C	$\text{idx}(T,C) = 4$
T	G	$\text{idx}(T,G) = 5$
T	N	$\text{idx}(T,N) = 1$

5

Figure 11

7/28

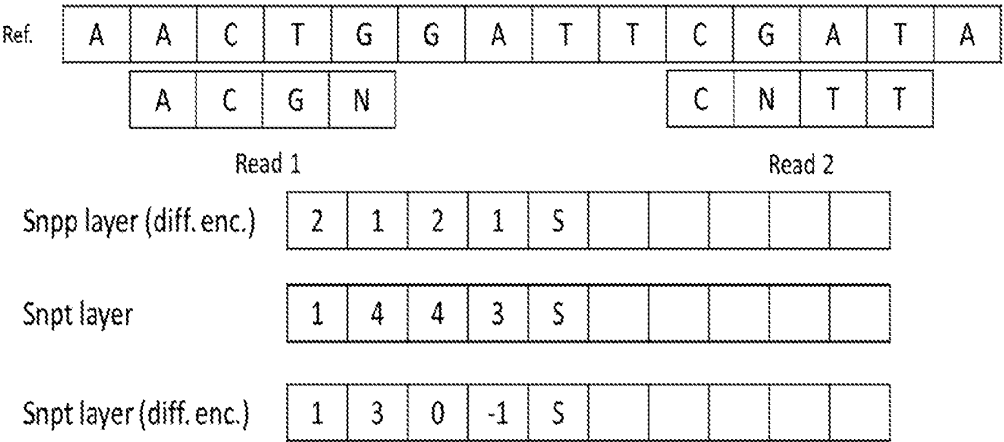


Figure 12

snpt Layer (with IUPAC codes)

A substitution type is calculated as index in a substitutions vector composed by all the possible symbols. For example:

S = [A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B]

Index direction

- ENCODING is from right to left
- DECODING from left to right

Ref	Read	Encoded Symb.
N	M	idx(N,M) = 2
N	W	idx(N,W) = 4
N	S	idx(N,S) = 5
N	B	idx(N,B) = 11

Ref	Read	Encoded Symb.
D	M	idx(D,M) = 8
A	Y	idx(A,Y) = 10
A	T	idx(A,T) = 3
A	N	idx(A,N) = 4
C	R	idx(C,R) = 6
C	G	idx(C,G) = 1
C	T	idx(C,T) = 2
C	W	idx(C,W) = 7
G	H	idx(G,H) = 11
G	C	idx(G,C) = 15
G	B	idx(G,B) = 13
G	N	idx(G,N) = 2
T	A	idx(T,A) = 13
T	M	idx(T,M) = 4
T	K	idx(T,K) = 8
T	V	idx(T,V) = 9

Figure 13

8/28

$S = [A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B]$

Direction

- ENCODING is from right to left
- DECODING from left to right

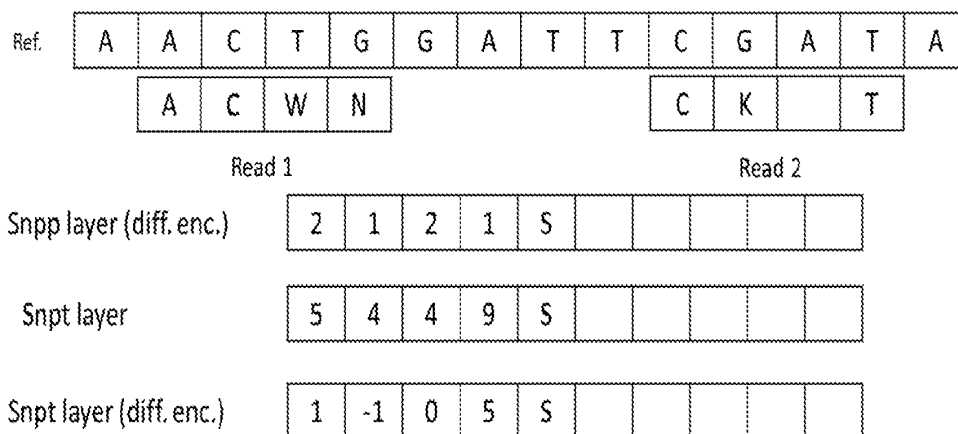


Figure 14

5

10

15

indt Layer (without IUPAC codes)

$$S = [A, C, G, T, N, Z]$$

Direction

- ENCODING is from right to left
- DECODING from left to right

Insert	Encoded Sym.
A	6
C	7
G	8
T	9
N	10

Ref.

A	A	C	T	G	G	A	T	T	C	G
---	---	---	---	---	---	---	---	---	---	---

T	C
---	---

A	C
---	---

T	G
---	---

C	N	T	T
---	---	---	---

Read 1 Read 2

indp layer (diff. enc.)

1	1	4	1	5					
---	---	---	---	---	--	--	--	--	--

indt layer

5	2	4	9	5					
---	---	---	---	---	--	--	--	--	--

indt layer (diff. enc.)

5	-3	2	5	5					
---	----	---	---	---	--	--	--	--	--

Figure 15

10/28

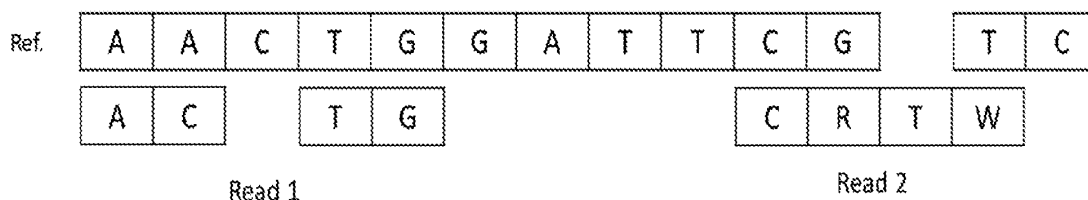
indt Layer (with IUPAC codes)

S = [A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B]

Direction

- ENCODING is from right to left
- DECODING from left to right

Insert	Encoded Sym.
A	16
C	17
G	18
T	19
N	20



indp layer (diff. enc.)

1	1	4	1	1	S				
---	---	---	---	---	---	--	--	--	--

indt layer

15	4	5	19	5	S				
----	---	---	----	---	---	--	--	--	--

indt layer (diff. enc.)

15	-11	1	18	-14	S				
----	-----	---	----	-----	---	--	--	--	--

Figure 16

11/28

Source model 2 (without IUPAC codes)

One position layer per substitution type, one per deletions and one per insertion type

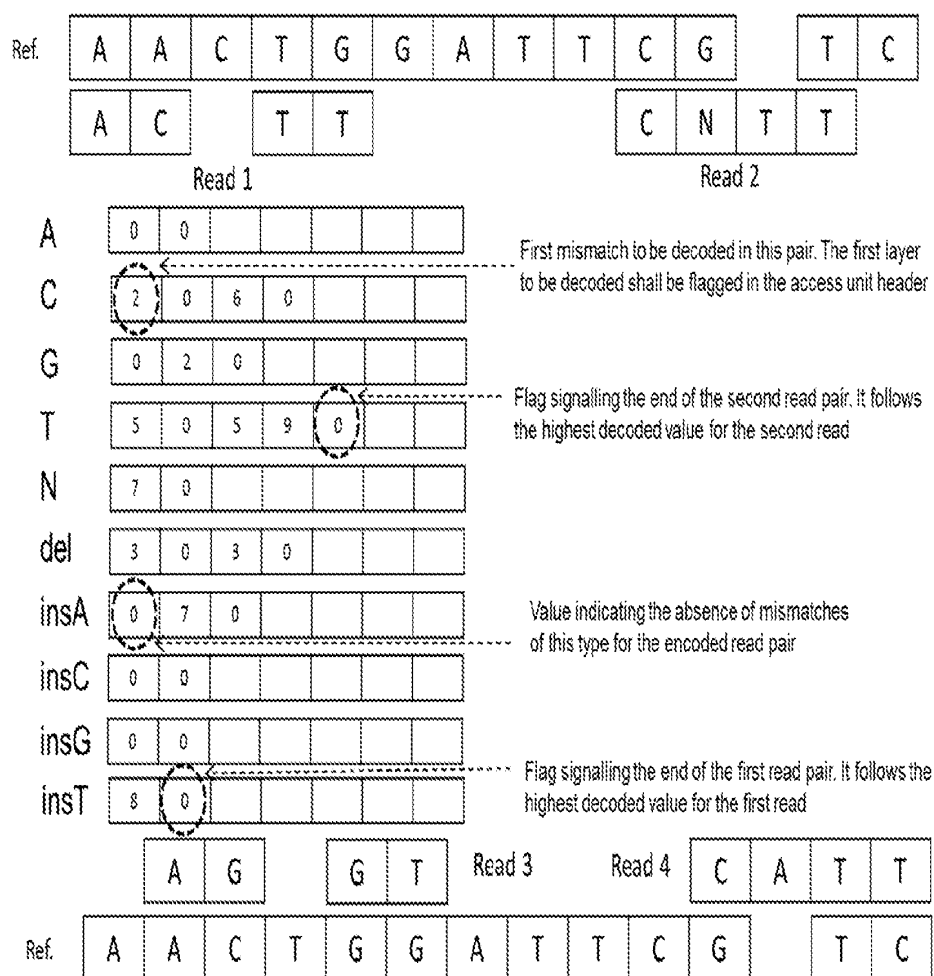


Figure 17

12/28

Source model 2 (without IUPAC codes)

One position layer per substitution type, one per deletions and one per insertion type

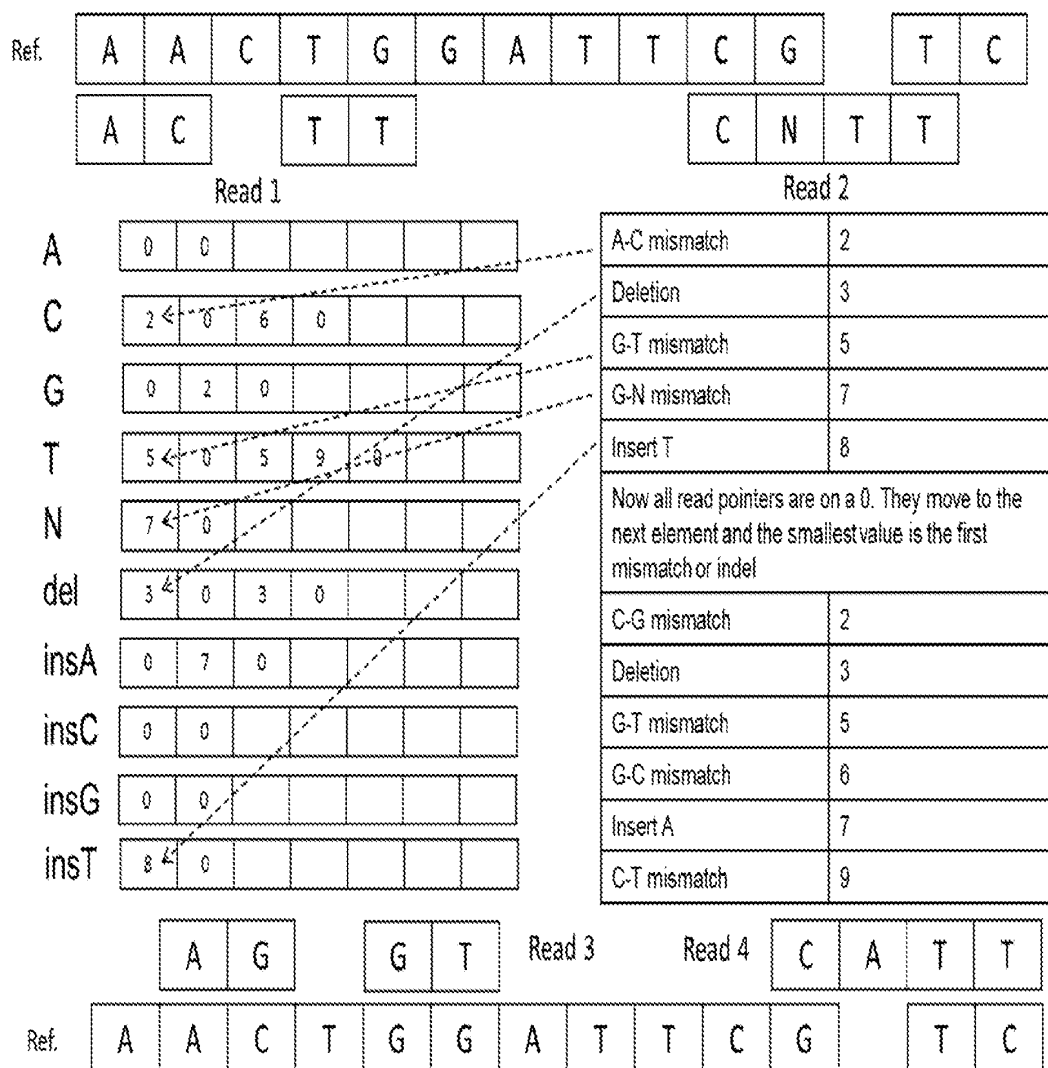


Figure 18

13/28

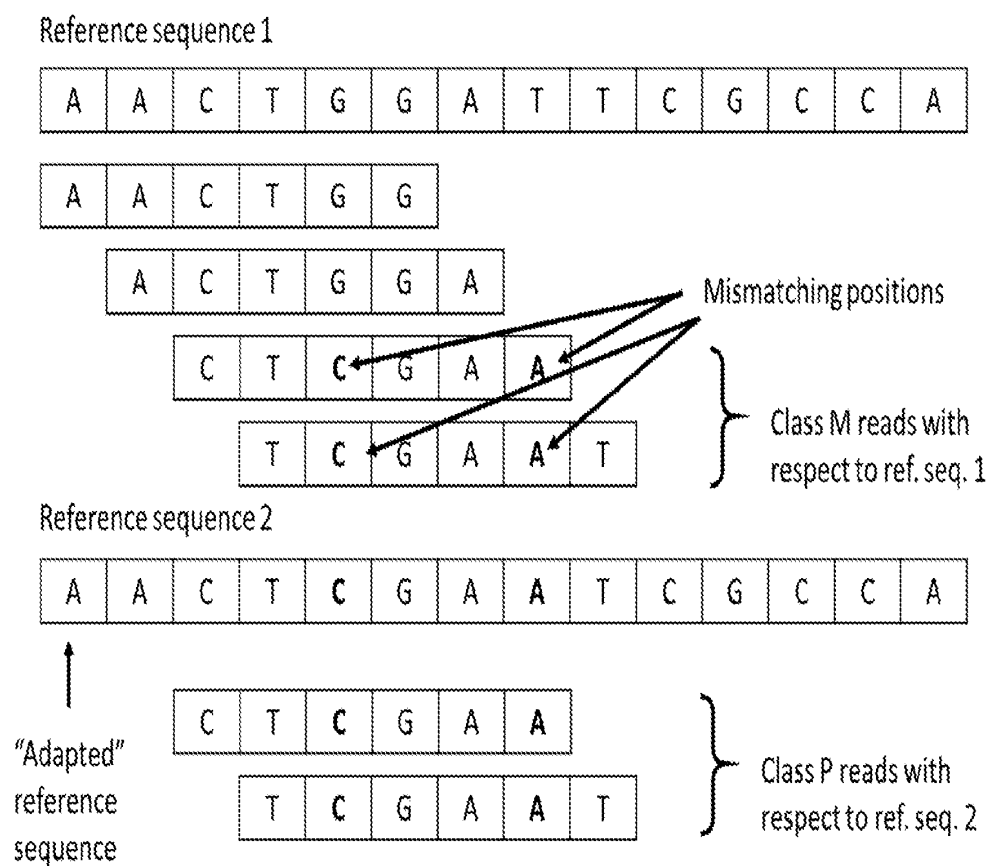


Figure 19

14/28

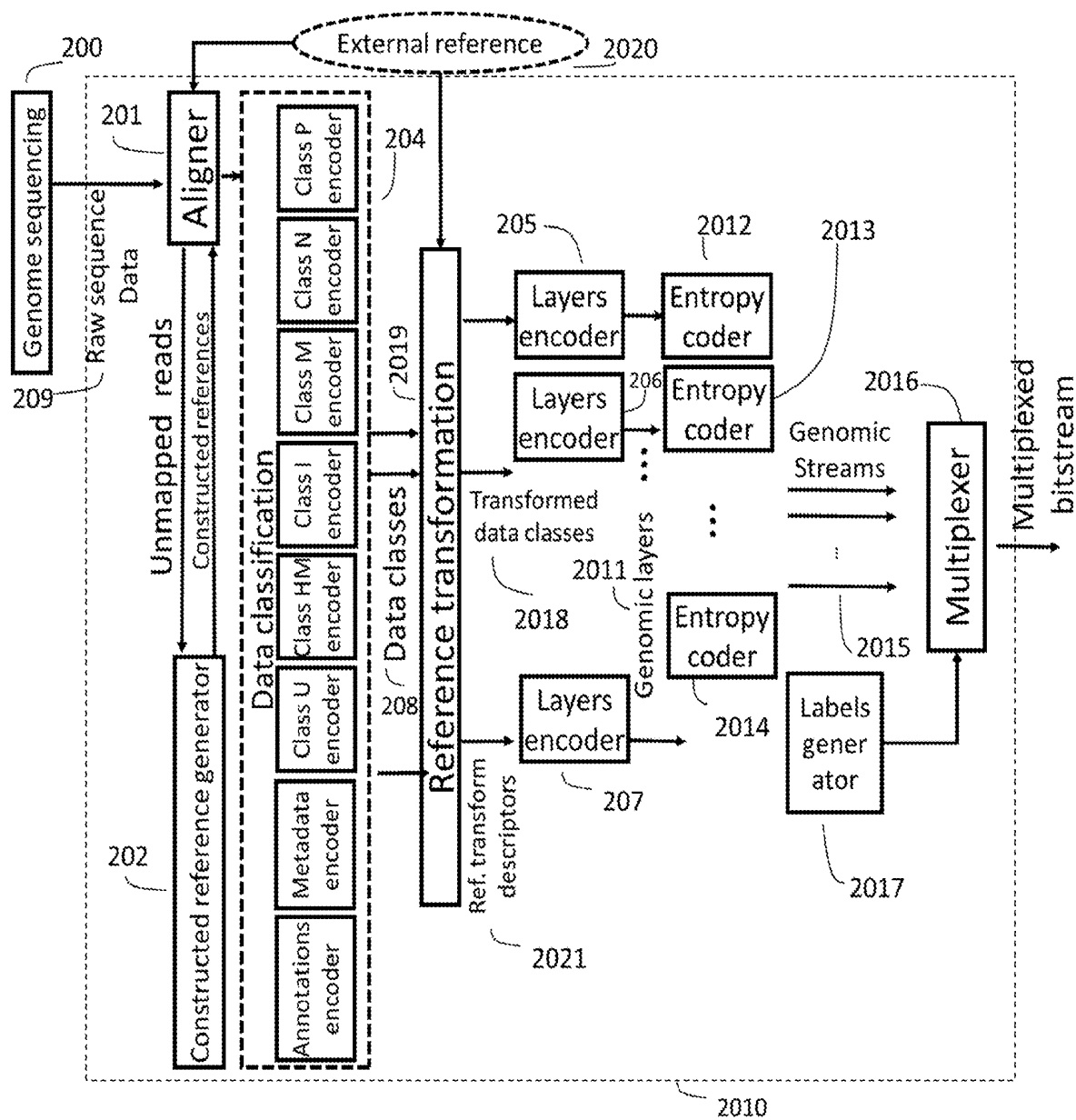


Figure 20

15/28

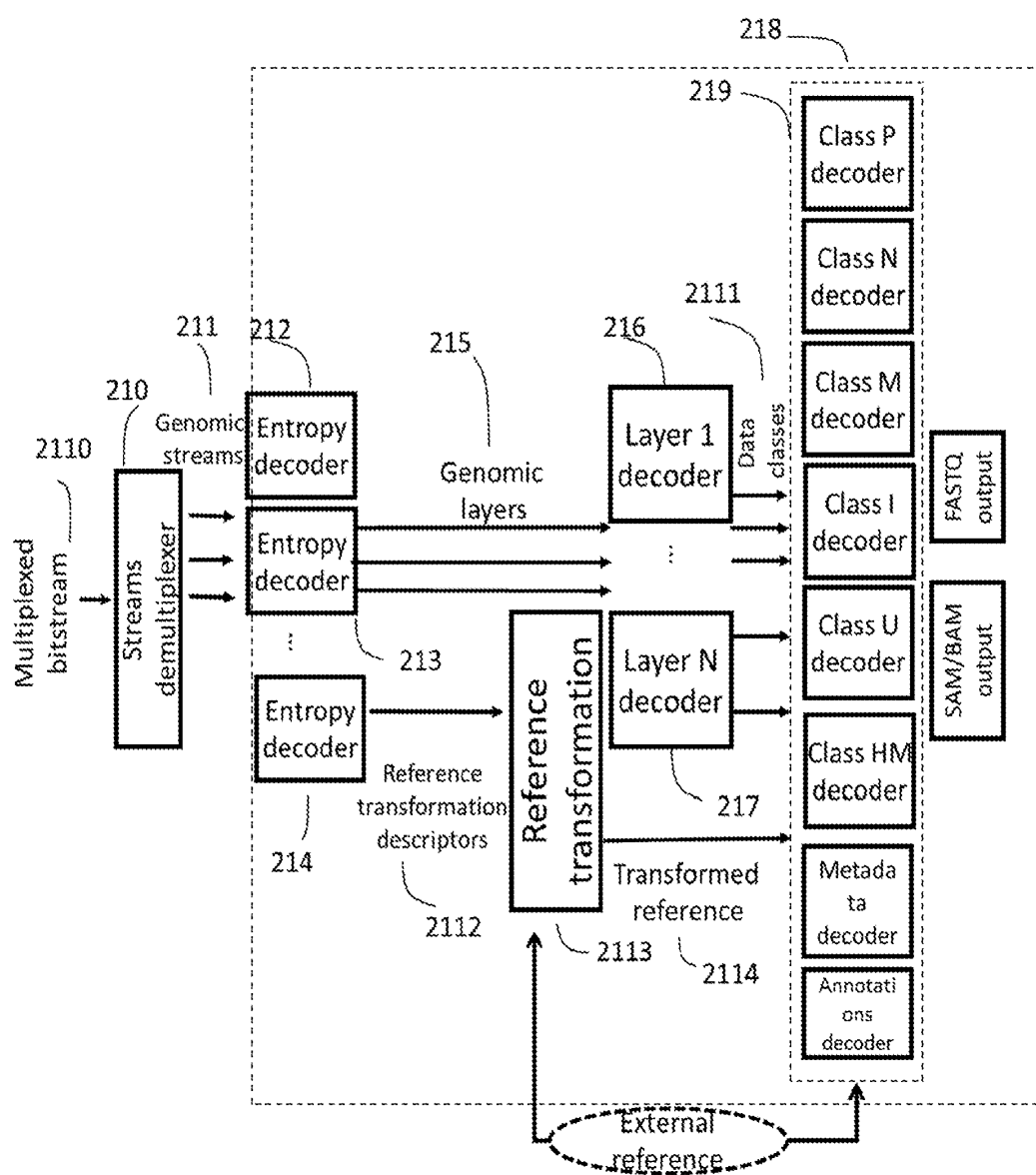


Figure 21

16/28

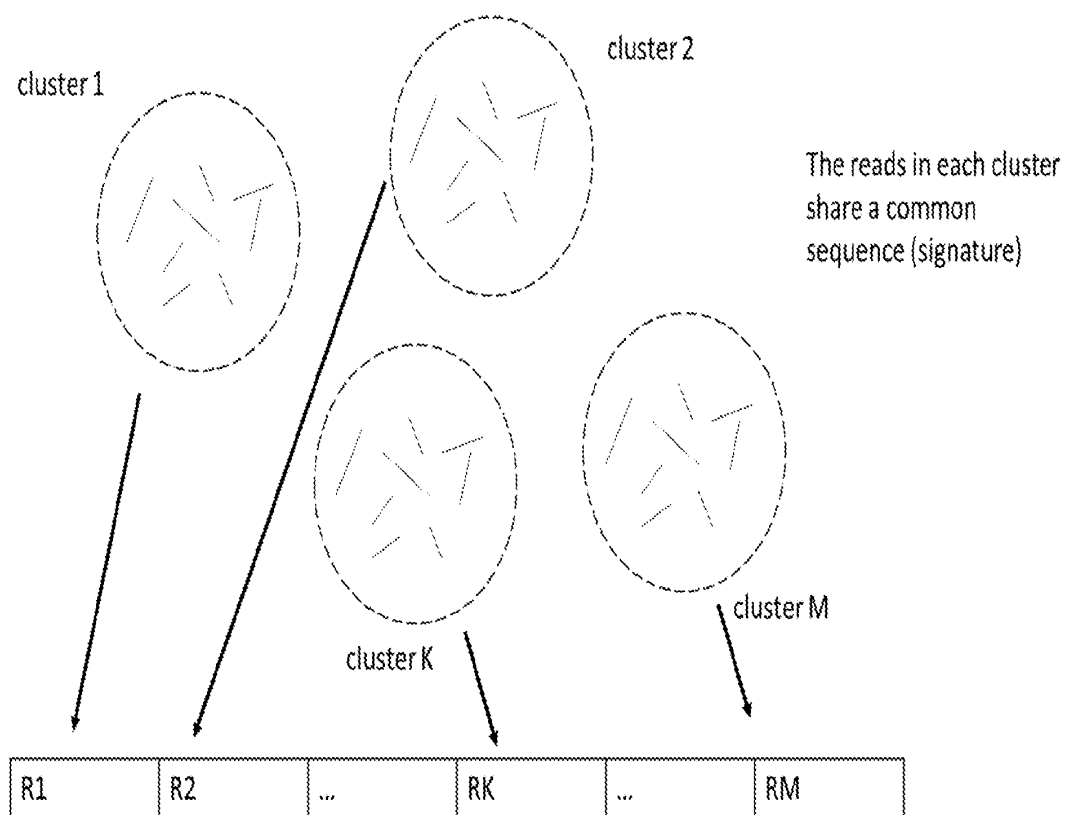
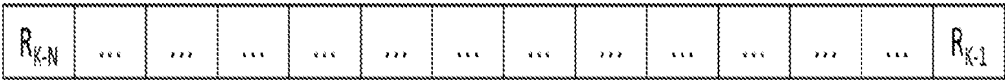


Figure 22

17/28



When encoding read K the reference is composed by reads K-1 to K-N
Appropriate padding is used when not enough reads are available.

Figure 23

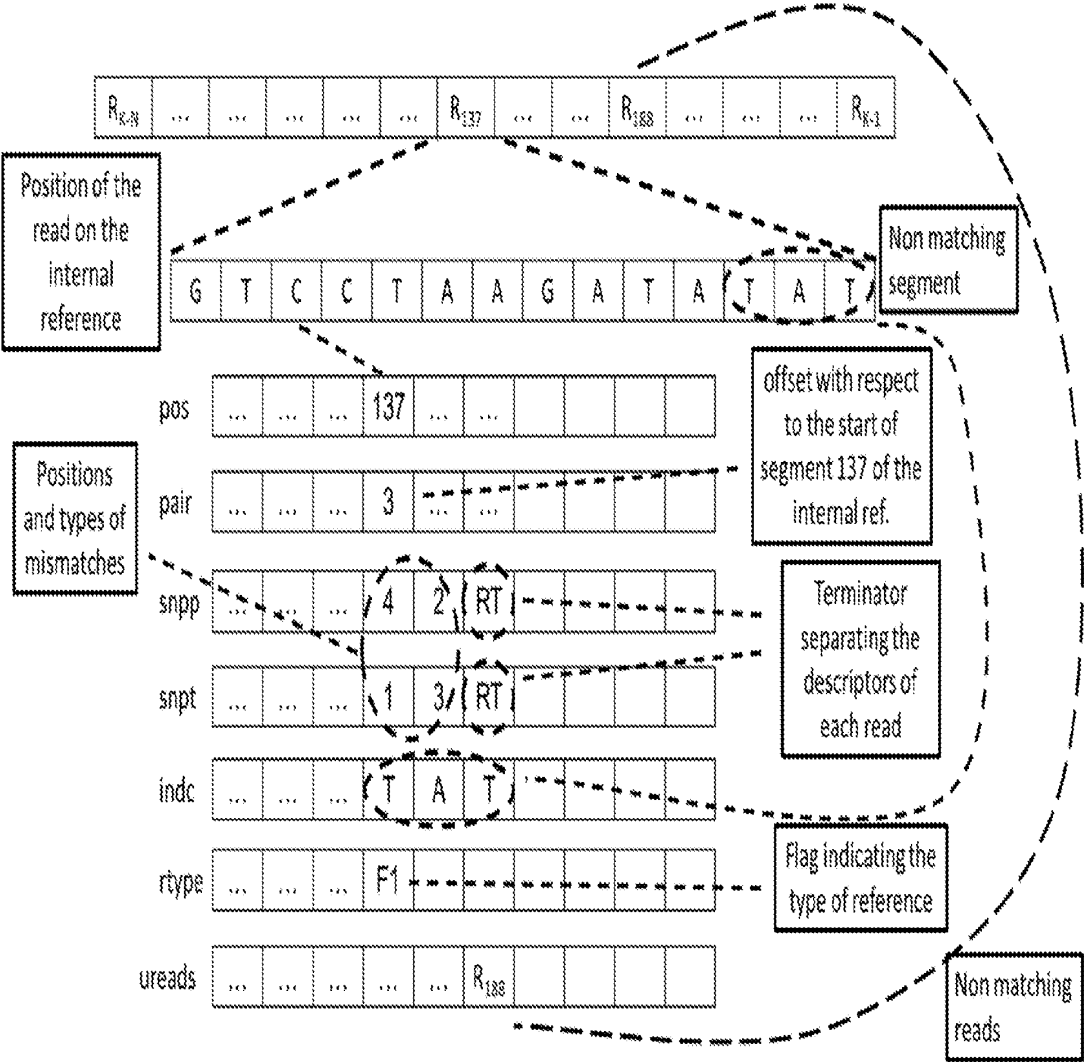


Figure 24

18/28

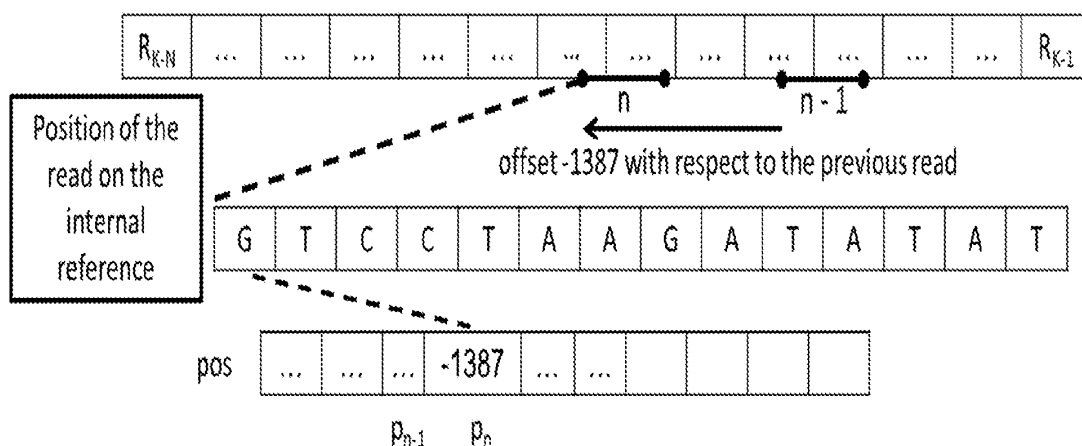
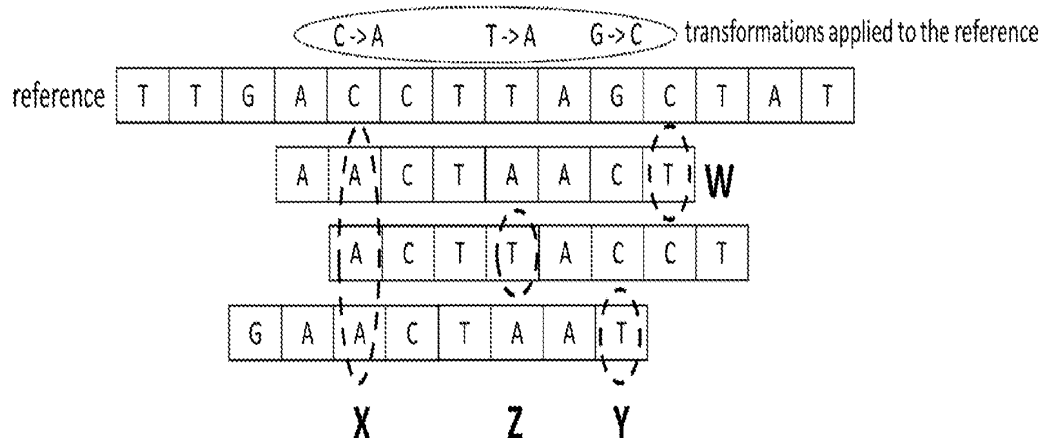


Figure 25



X = mismatch that will disappear

Y = mismatch that will not disappear but mismatch code shall change

W = mismatch that stays unchanged

Z = Newly introduced mismatch after transformation

19/28

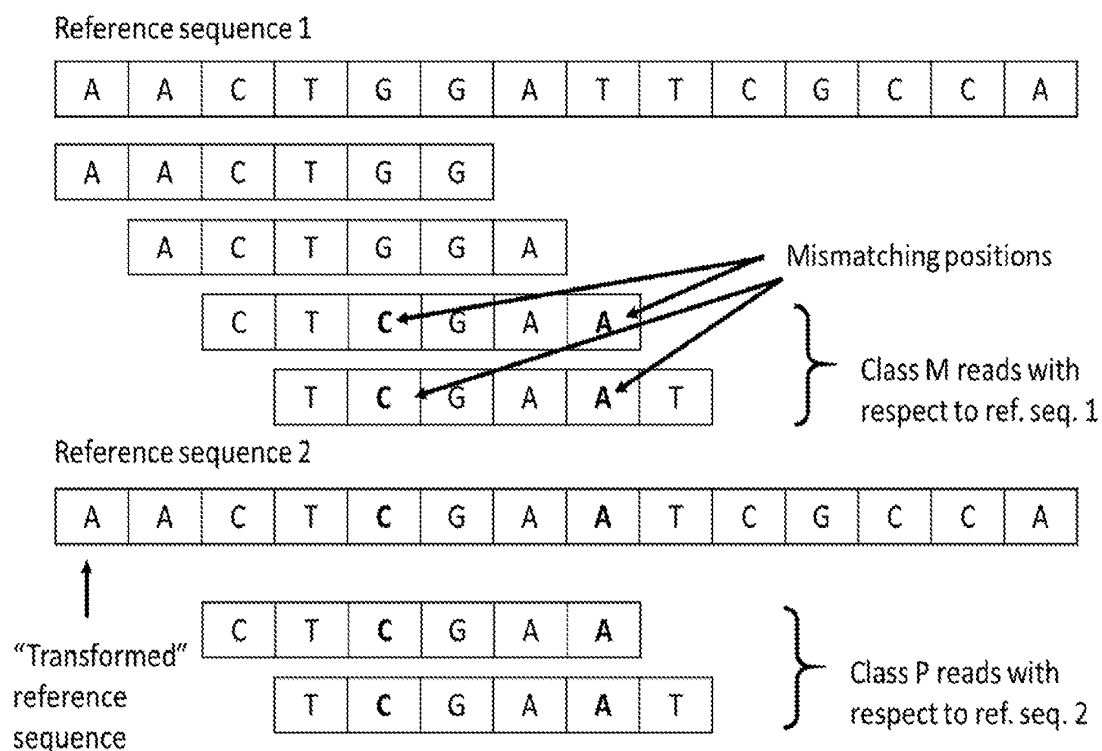


Figure 27

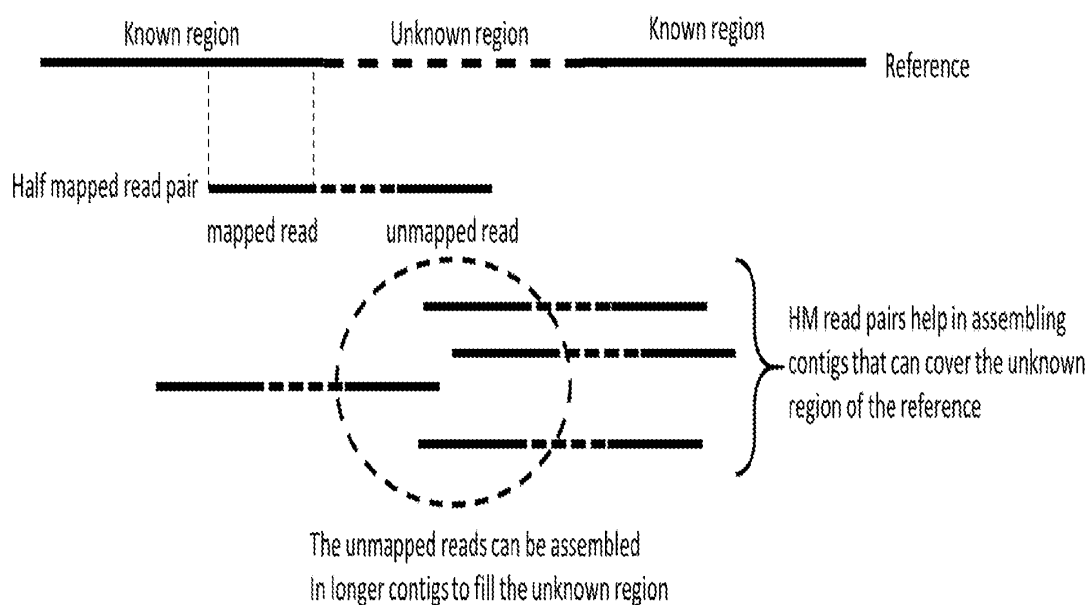


Figure 28

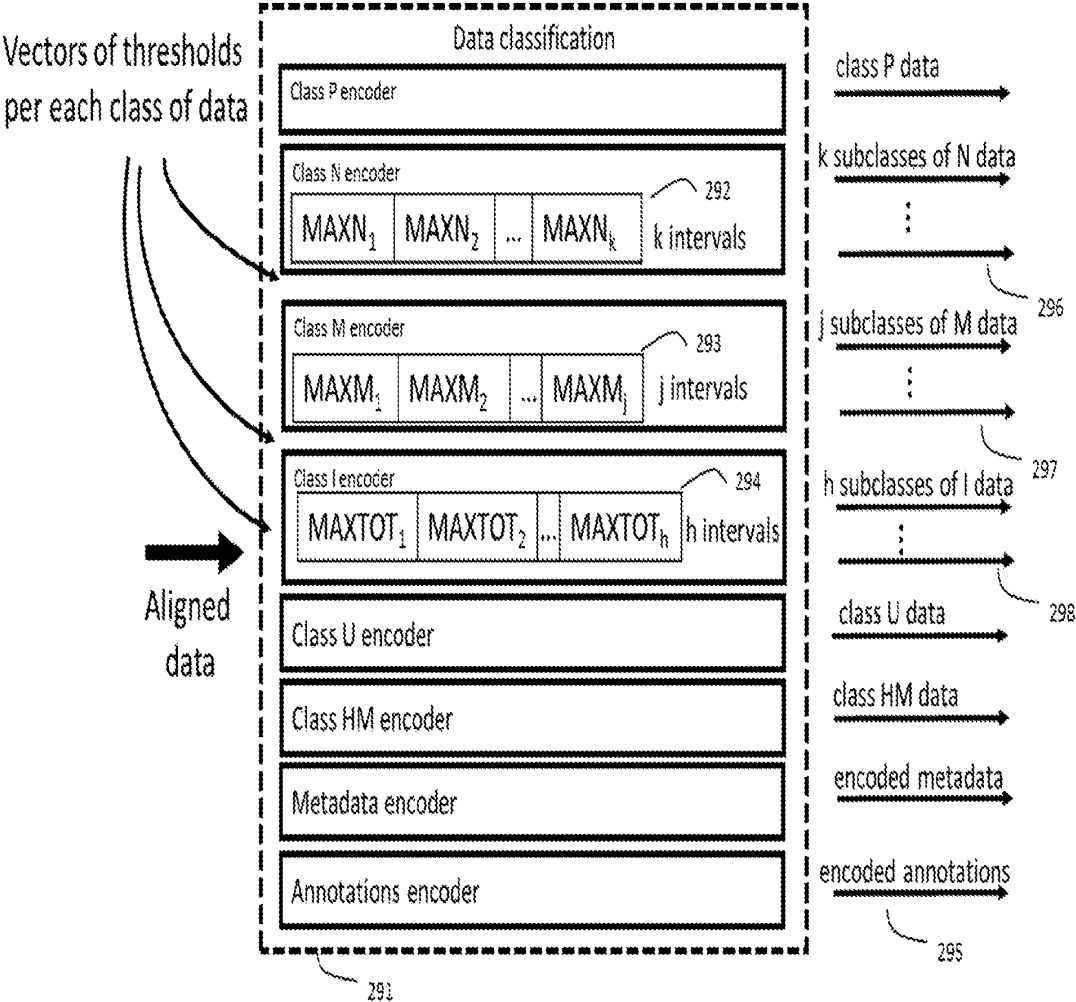
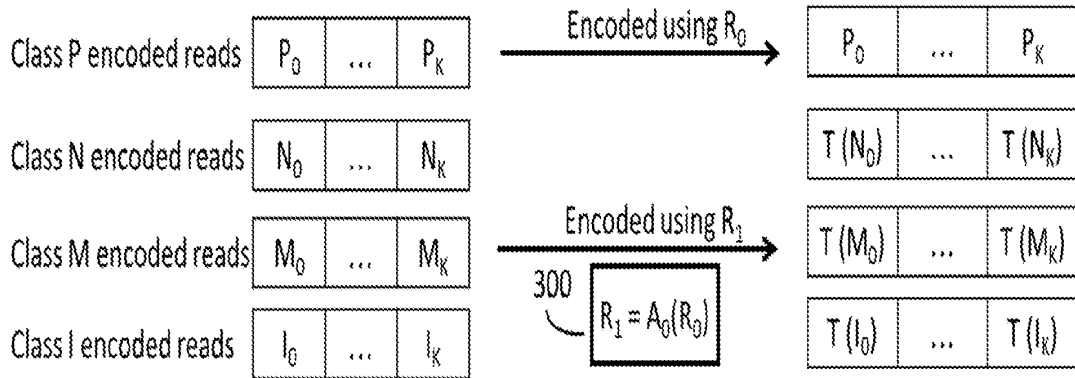


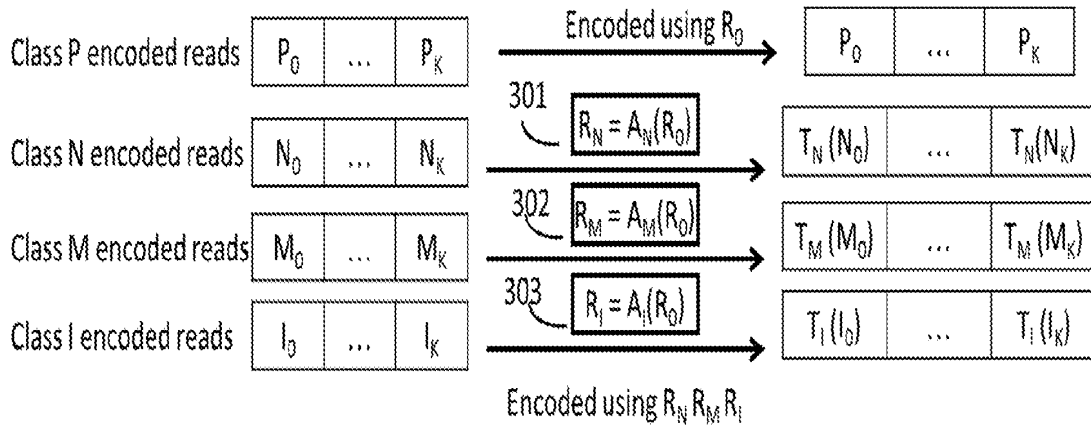
Figure 29

21/28

Classes N, M and I can use the same reference transformation A_0 or use different transformations A_N, A_M, A_I



T represents the transformation of each encoded read when encoded with respect to $R_1 = A_0(R_0)$



T is the transformation of each encoded read when encoded with respect to $R_1 = A_0(R_0)$

Figure 30

22/28

200	201	202	203	204	205	206	207	208	209	210
Unique ID	Version	Header Size	Reads Length	Ref count	Ref ID List	N. AUs per ref.	Label	Data Type	M.I.T.	Params Set

Figure 31

Master Index Table

Type	1	2	3	...	N
Class P ptr					
Class N ptr					
Class M ptr					
Class I ptr					
Class U ptr					
Class HM ptr					
Metadata					
Annotations					

Access Units mapping positions

Position on the reference sequence of the first read in each AU for each class

These pointers are used to access the LIT and find the pointer used to "jump" to physical positions in the reference sequence

Figure 32

23/28

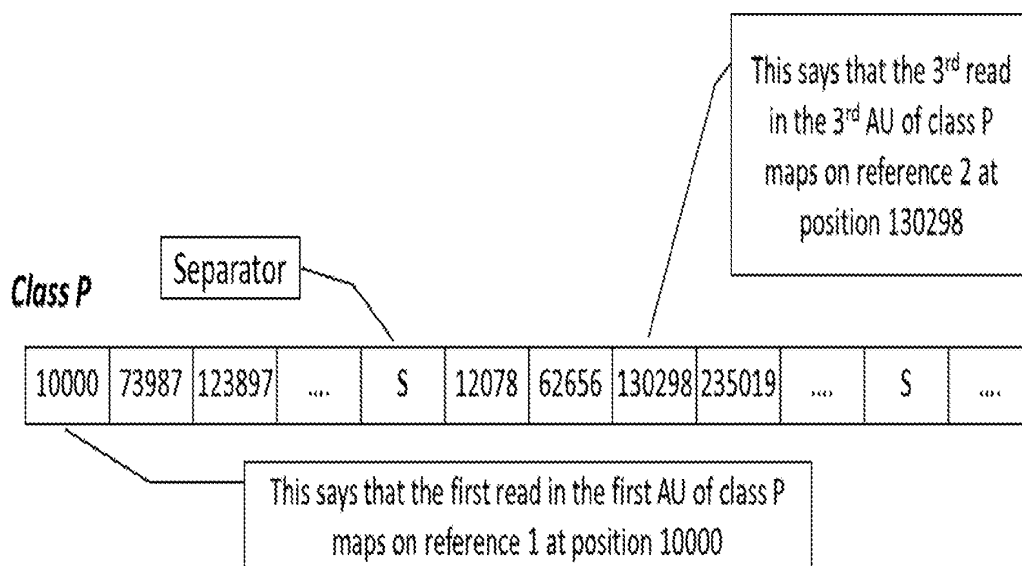


Figure 33

5

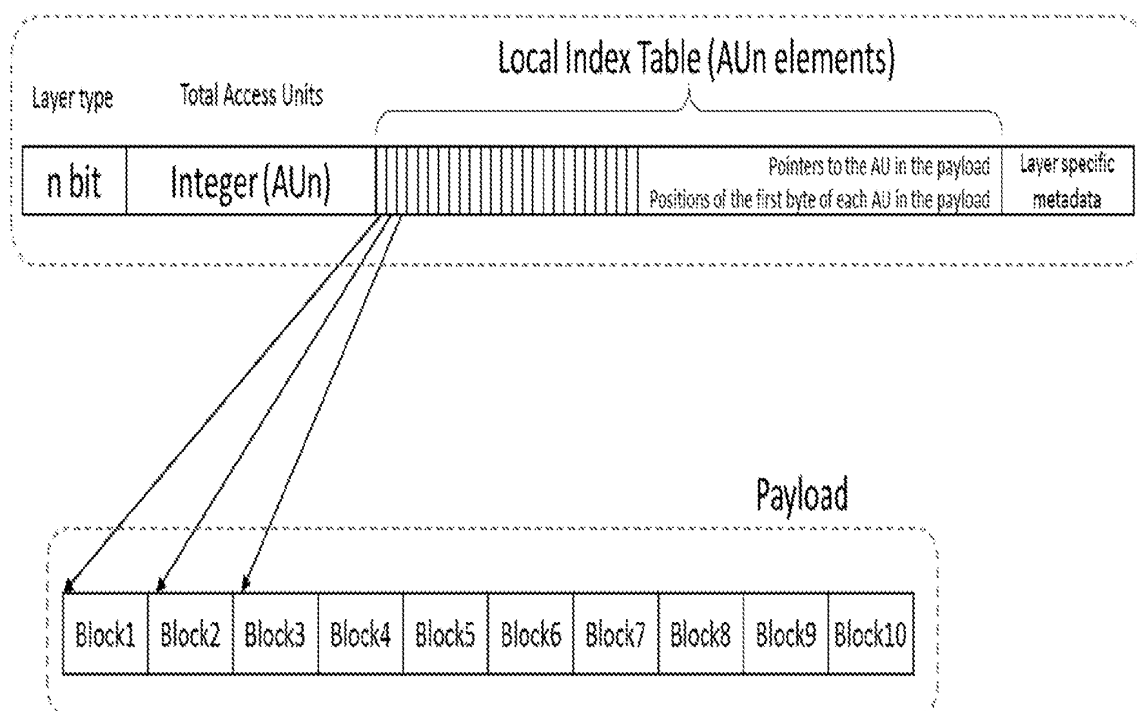


Figure 34

24/28

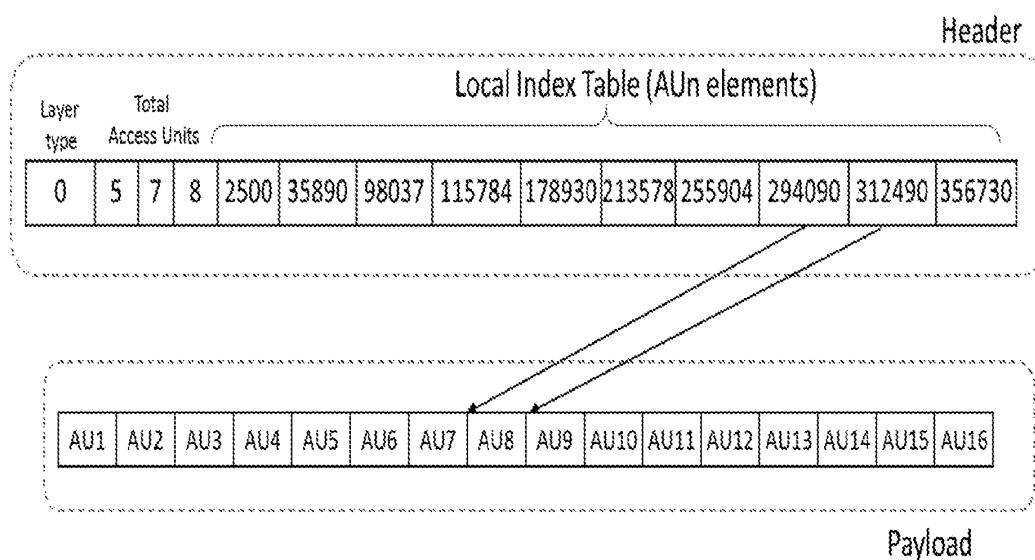


Figure 35

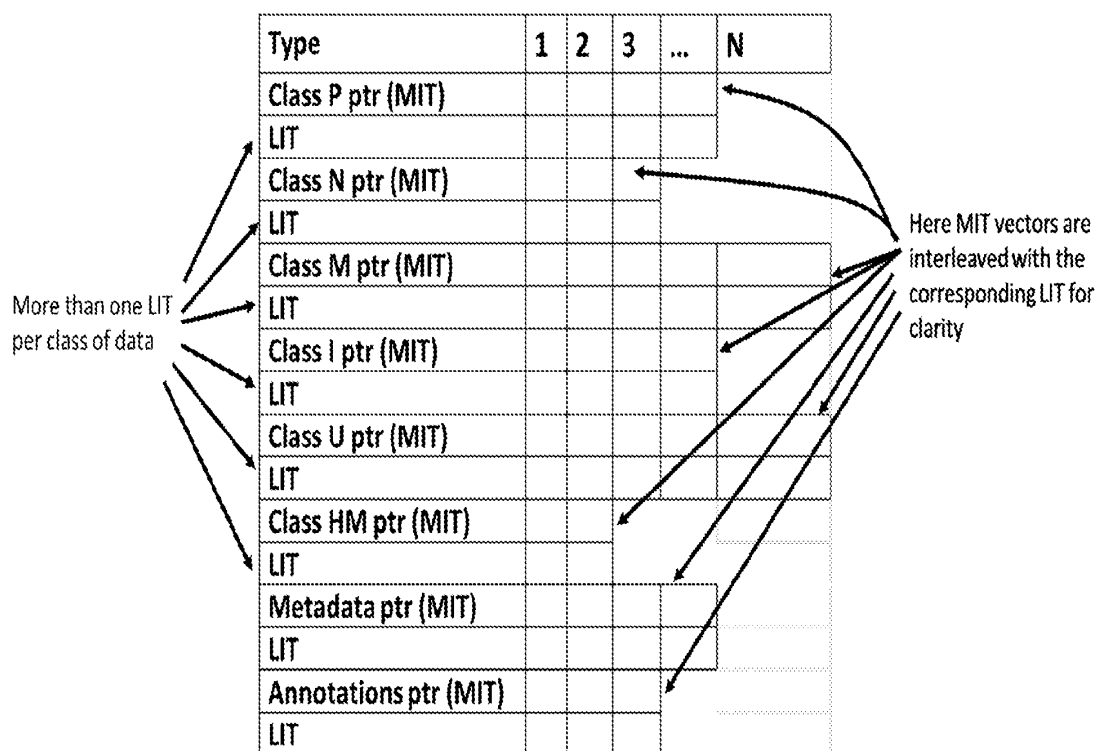


Figure 36

25/28

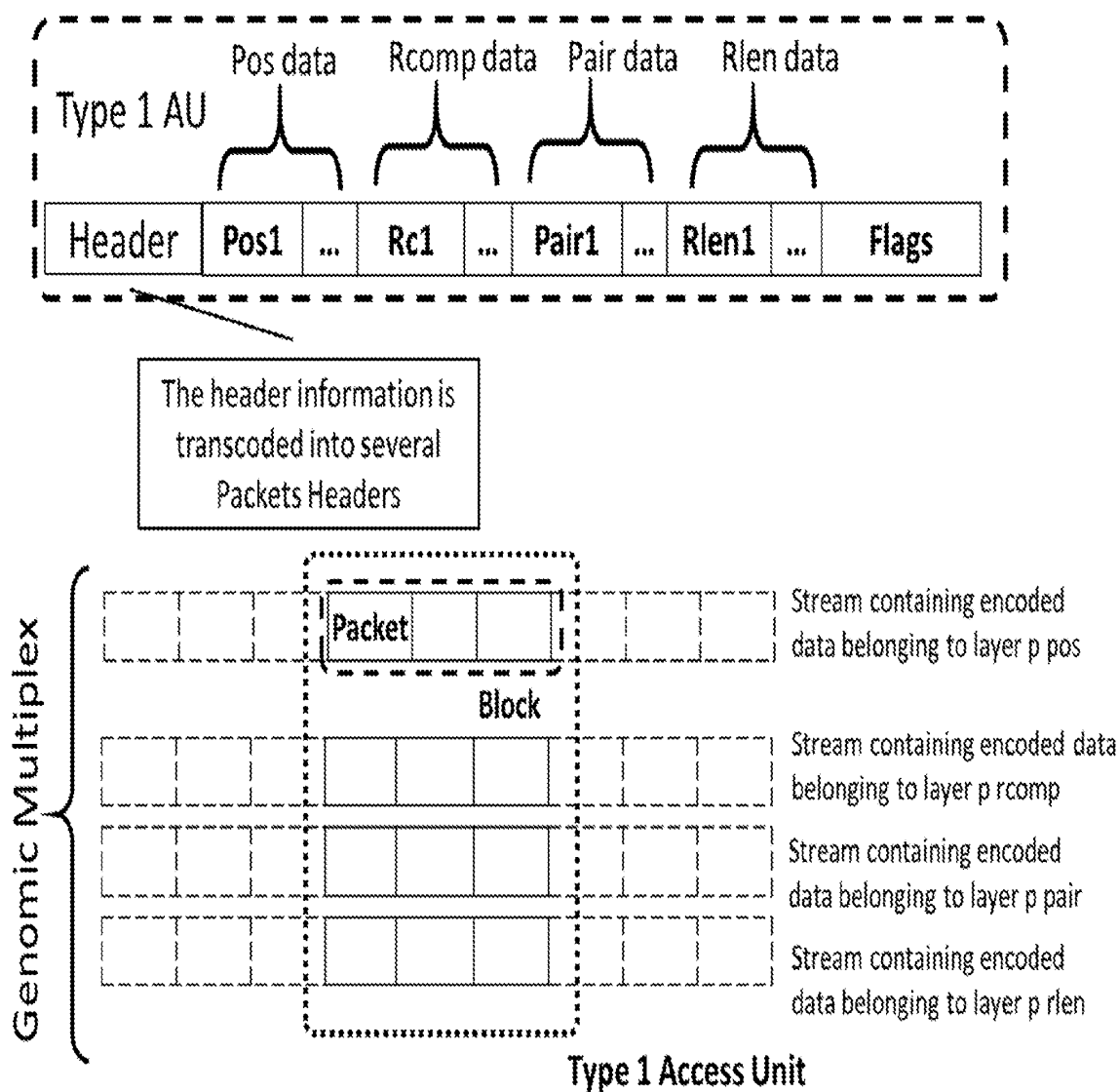


Figure 37

26/28

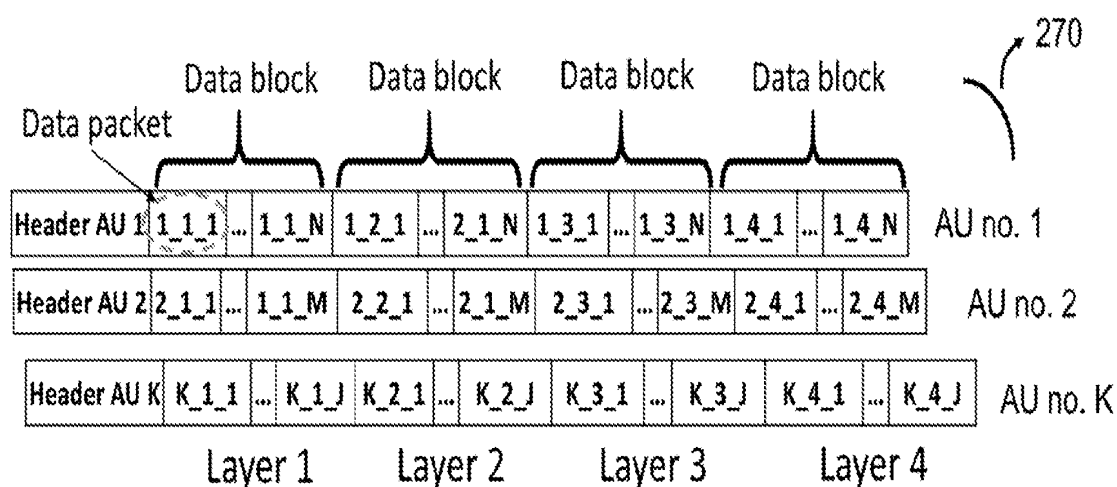


Figure 38

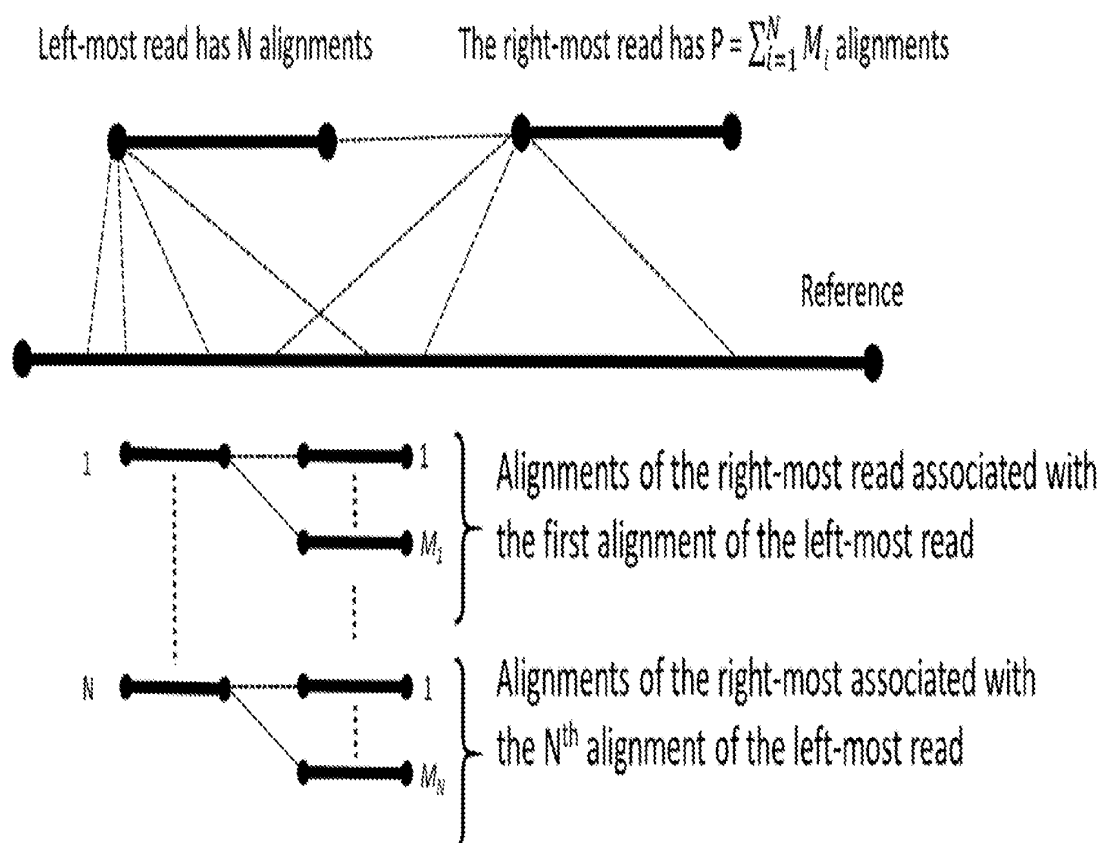


Figure 39

27/28

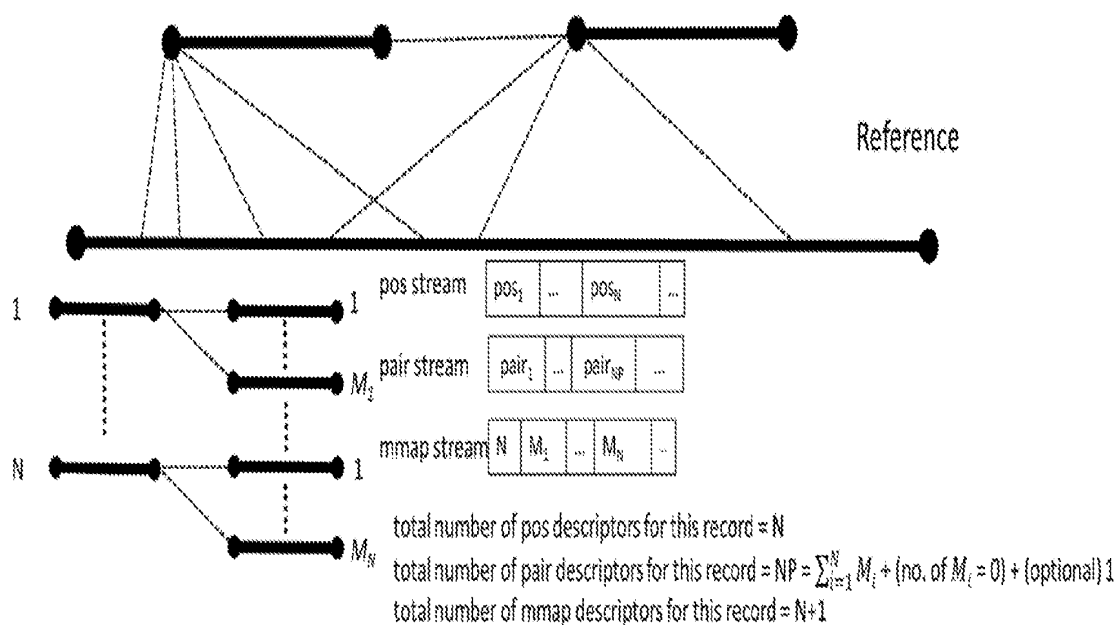


Figure 40

5

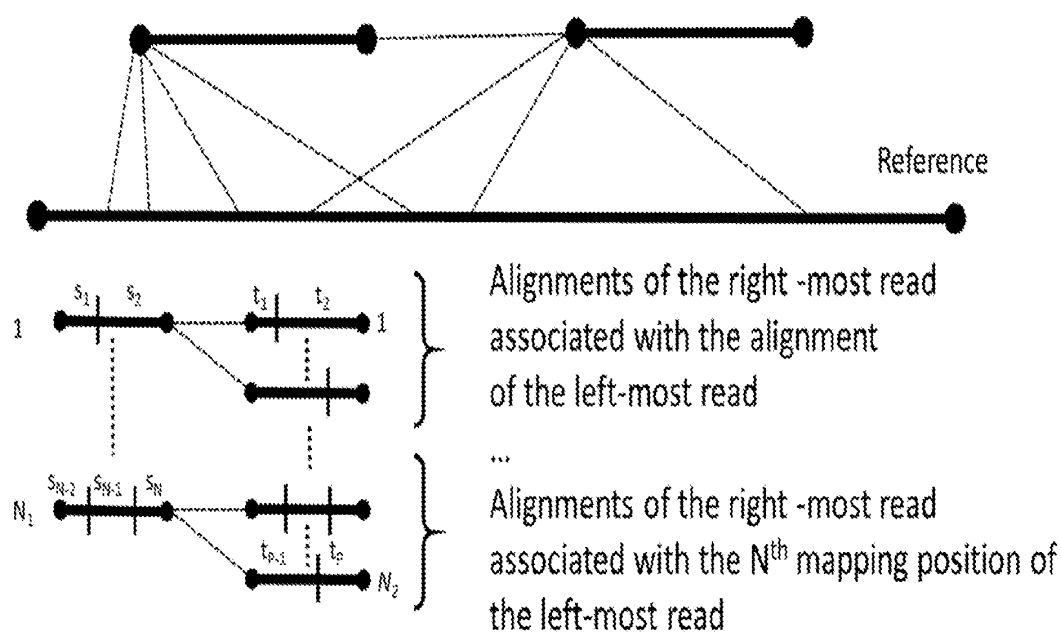


Figure 41

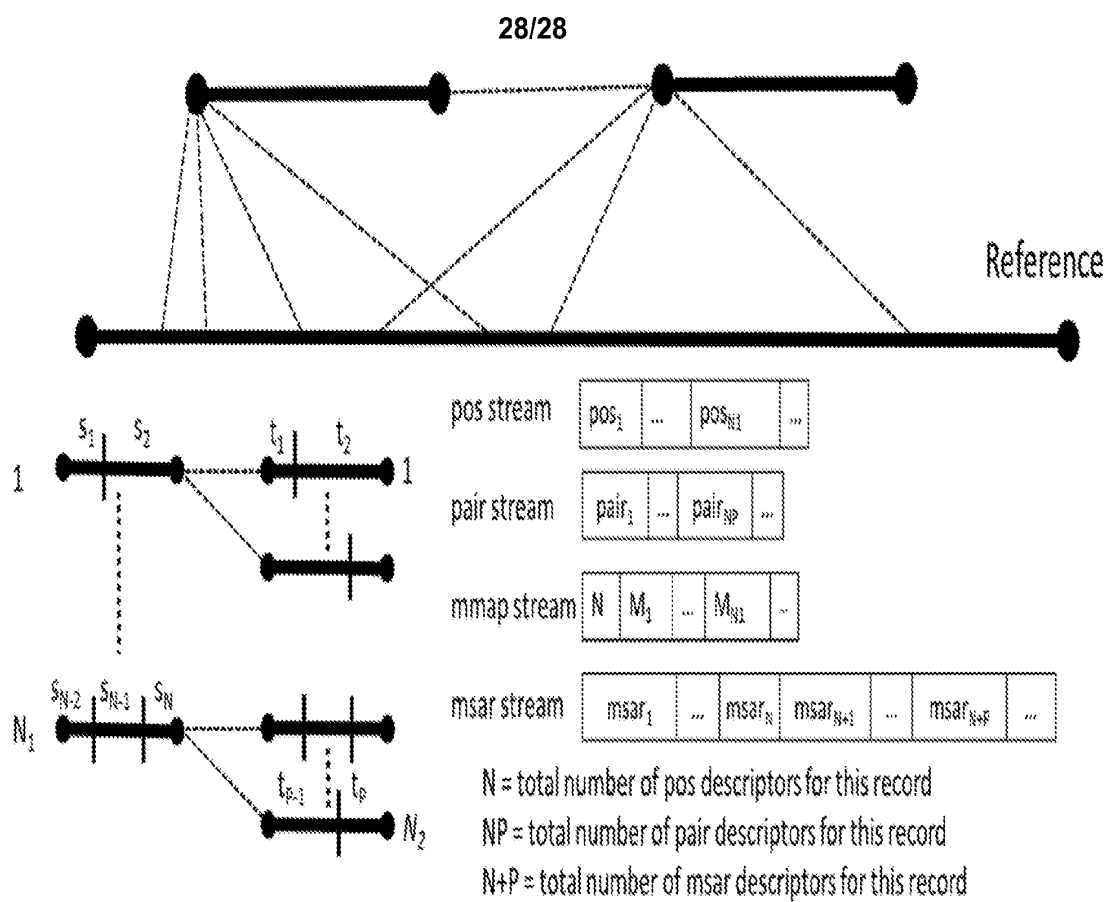


Figure 42