1.3.1 The reference GFA format (rGFA). In an ordinary GFA, a vertex is a sequence typically derived from a reference or a contig. The origin of the vertex is not recorded and as a result, we only see the *vertex coordinate* (where a position is denoted by "vertexID:offset"), which is unstable and decoupled from linear annotations.

To resolve this issue, we propose rGFA, an extension to GFA with three additional tags that indicate the origin of a vertex (Fig. 8). This simple addition gives us a unique *stable coordinate* system as an extension to the linear reference coordinate (e.g. GRCh38). We can pinpoint a position such as "chr1:9" in the graph and map existing annotations onto the graph. We can also report a path or walk in the stable coordinate. For example, path "v1→v2→v3→v4" corresponds to "chr1:0-17" and path "v1→v2→v5→v6" corresponds to "chr1:0-8→foo:8-16".

```
S v1 CTGAA SN:Z:chr1 SS:i:0  SR:i:0
S v2 ACG   SN:Z:chr1 SS:i:5  SR:i:0
S v3 TGGC  SN:Z:chr1 SS:i:8  SR:i:0
S v4 TGTGA SN:Z:chr1 SS:i:12 SR:i:0
S v5 TTTC  SN:Z:foo  SS:i:8  SR:i:1
S v6 CTGA  SN:Z:foo  SS:i:12 SR:i:1
S v7 GTTAC SN:Z:bar  SS:i:5  SR:i:2
L v1 + v2 +
L v2 + v3 +
L v3 + v4 +
L v2 + v5 +
L v5 + v6 +
L v6 + v4 +
L v1 + v7 +
L v7 + v6 +
W foo >v1>v2>v5>v6>v4 SM:Z:Bob
W bar >v1>v7>v6>v4    SM:Z:Mary
```

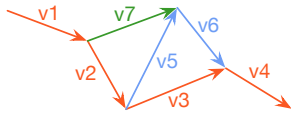

Fig. 8: Example reference GFA and the corresponding edge graph. SN: sequence name from which the vertex is derived. SS: starting position. SR: rank of the vertex: 0 for a vertex on the linear reference; >0 for non-reference.

We can further extend the baseline rGFA model by embedding haplotypes with walk records (*W*-lines in Fig. 8). This keeps long-range information and enables haplotype-aware mapping[40] that avoids traversing paths absent from the input haplotypes, and thus reduces mapping ambiguity and search space[41] for a graph composed of millions of small variants.

1.3.2 The pairwise Graph Alignment Format (GAF). GAF is a simple TAB-delimited text format consisting of 12 mandatory fields to represent reads aligned to a reference pan-genome graph:

- Col 1–4: query name, length, start and end positions.
- Col 5: the query strand relative to the mapping path in col 6.
- Col 6: a stable graph path in the format matching the regular expression `/([><][^\s><]+:\d+-\d+)+|([^\s><]+)/`. The less or greater sign indicates the orientation of the segment. If a path consists of a single segment, coordinates can be omitted and its orientation is assumed to be forward ">".
- Col 7–9: total path length, start and end mapping positions on the forward strand of the path.
- Col 10–12: number of matching bases, alignment block length and mapping quality

Each line may have tags as in the SAM format. CIGARs are stored in tags (Fig. 9). For a linear reference genome, GAF reduces to the PAF format which we designed for minimap[31]. As with rGFA, the simplicity of the GAF specification makes it easy to implement and increases its chance of being broadly adopted.

```
read1  6  0  6  +  chr1               17  7 13  6  6  60  cg:Z:6M
read2  7  0  7  +  >chr1:5-8>foo:8-16  11  1  8  7  7  60  cg:Z:7M
```

Fig. 9: Example GAF for reads "GTGGCT" and "CGTTTCC" mapped to Fig. 8.