Self-contained VCF format Optimized with 2 bit Encoding

Jay Jung
Rice University

***Abstract*—The Variant Call Format (VCF) is a verbose and text-based format that is optimized for human rather than machine readability, making it inefficient for large-scale data storage and machine processing. Moreover, reconstructing personal genome data requires an external FASTA file, which further contributes to storage inefficiency. This paper outlines a type-specific binary format that significantly reduces storage size while preserving full restoration capability. The suggested 2-bit encoded format contains reference sequence, mask bit for distinguishing 'N', variants-including SNP, insertion, and deletion-and corresponding metadata(AF,DP,GT) as compact 2-bit encoded format. This method achieves approximately 87 percent of compression and enables reconstruction of personal genomic data without relying on the original FASTA file. This design offers substantial benefits for simulation-based genomic studies by improving disk space efficiency.**

I. Introduction

Simulations of genomic sequences generate large volumes of data, much of which are highly repetitive. To address this, researchers have relied on compressed representations of genomic data that distill the differences between a reference genome and an alternate genome. In studies of genomic data such as SARS-CoV-2 [1] and human genomes [2], the Variant Call Format (VCF) is one of the most commonly used formats to represent differences.

However, VCF is a verbose format that is optimized for human readability rather than machine efficiency, making it not optimal for large-scale data storage and data processing. A standardized and efficient way to compress simulation output while allowing on-demand decompression would be highly beneficial.

VCF has two major limitations. First, as it is primarily designed for human readability, it is not optimized for data processing or storage efficiency. Second, VCF files only contain variant information and cannot be used to reconstruct a complete genome without a corresponding reference FASTA file. This introduces additional space overhead, as both the verbose VCF file and the large FASTA file must be stored together.

For example, the combination of the HG00157 chromosome 11 VCF file and its corresponding FASTA file occupies approximately 573MB. Given that this size corresponds to just a single chromosome, storing and analyzing an entire human genome—or the complete genome of any large organism—becomes highly inefficient in terms of storage space.

To address the limitations of conventional VCF + FASTA format, this paper outlines a type-specific binary format representing a genomic variation. This format not only includes genomic variation, but also 2-bit encoded reference genomic data. This format is type-specific and uses fixed-length formats for three types of variation, Single nucleotide polymorphisms, insertions, and deletions, using a 2-bit encoding scheme. Each variation also includes essential metadata, such as allele frequency (AF), read depth (DP), and genotype (GT), which are also encoded with the suggested compact 2-bit form.

Additionally, this design incorporates a 1-bit mask block indicating positions of unknown bases (e.g., 'N') in the reference, allowing accurate restoration of the original genome including masked sites.

This binary format achieves approximately 87% compression compared to traditional VCF + FASTA representations and supports the restoration of genomic sequences without relying on the FASTA file. By eliminating redundancy and enabling on-demand decompression, this method significantly reduces storage requirements, especially in large-scale or simulation-based genomic analysis. This paper outlines a novel format of recording variants that enables restoration of genomic data without relying on FASTA file.

Our main contributions are threefold: 1) We design a type-specific binary format that encodes SNPs, insertions, and deletions using a compact 2-bit representation; 2) We introduce a compact metadata structure that encodes allele frequency (AF), read depth (DP), and genotype (GT) in a fixed-size binary layout; 3) We demonstrate that our format achieves approximately 87% compression over traditional VCF + FASTA representations, while supporting complete genome reconstruction without requiring a reference FASTA file.

II. Method

A. Overview of the 2-bit Encoded VCF Format

We designed a type-specific, fixed-length binary format that stores both variant information and the full reference genome sequence. This format enables complete reconstruction of an individual genome without relying on an external FASTA file, as the reference sequence is stored at the beginning of the file using a 2-bit encoding scheme.

The resulting binary file consists of four components: the full 2-bit encoded reference sequence, masking block indicating the location of unknown nucleotide, variant records (including SNPs, insertions, and deletions), and corresponding metadata fields (allele frequency, read depth, and genotype).

B. Binary format construction

To construct the self-contained binary VCF file, we first encode the entire reference genome using a 2-bit representation, storing it at the beginning of the file. A separate mask block is then generated to indicate the positions of unknown bases ('N') in the reference sequence.

Next, we extract variant information from a standard VCF file. Each variant is classified as a SNP, insertion, or deletion, and encoded using a type-specific fixed-length binary format. For insertions and deletions, both altered sequences and reference sequences are encoded in 2-bit format.

For each variant, metadata fields—allele frequency (AF), read depth (DP), and genotype (GT)—are extracted and encoded as compact binary format. These metadata values are appended immediately after the corresponding variant records.

All components are concatenated in a fixed order: reference block, mask block, variant block, and metadata block. This structured layout ensures that the binary file is both compact and fully self-contained, allowing for direct decoding and full restoration of the genome without external resources.

C. Structure of Suggested Idea

Section	Description
Reference Block	2-bit encoded ACTG-only reference genome sequence
Mask Block	1-bit mask for N positions (1 = N, 0 = ACTG)
Variant Block	Encoded SNPs, insertions, and deletions
Metadata Block	`META` + length + [AF, DP, GT] per variant

D. Reference and Mask Encoding

The reference genome sequence is stored at the beginning of the HEX file using a 2-bit representation, where A = `00`, C = `01`, G = `10`, and T = `11`. To preserve unknown entries, such as 'N', a corresponding bit-mask is included. This mask uses one bit per base to indicate whether the entry is unknown: a bit value of 1 marks the position as 'N', while 0 indicates a standard ACTG base.

The reference and mask blocks are packed separately—4 bases per byte for the sequence and 8 bits per byte for the mask. This structure enables accurate genome reconstruction including unknown positions, while maintaining high compression efficiency.

E. Variant Encoding Structure

Each variant is encoded as a fixed-length binary record, beginning with a 1-byte type flag that indicates the variant type:

- `0x00`: SNP (Single Nucleotide Polymorphism)
- `0x01`: Insertion
- `0x02`: Deletion

The detailed structure of each record is as follows:

- **SNP (7 bytes)**:
  - 1 byte: Type flag (0x00)
  - 4 bytes: Position (POS)
  - 1 byte: Reference base (REF)
  - 1 byte: Alternate base (ALT)
- **Insertion (23 bytes)**:
  - 1 byte: Type flag (0x01)
  - 4 bytes: Position (POS)
  - 2 bytes: Length of inserted sequence (LEN)
  - 16 bytes: Inserted sequence (ALT), encoded using 2-bit representation
- **Deletion (23 bytes)**:
  - 1 byte: Type flag (0x02)
  - 4 bytes: Position (POS)
  - 2 bytes: Length of deleted sequence (LEN)
  - 16 bytes: Deleted sequence (REF), encoded using 2-bit representation

F. Metadata Encoding

Each variant is mapped one-to-one with the metadata in the order they appear. This 3-byte metadata information is appended at the end of the HEX file. The metadata block is preceded by a 4-byte `META` marker and a 4-byte unsigned integer indicating the length of the block. The structure is as follows:

- 1 byte: Allele Frequency (AF), float in [0, 1] scaled to 0–255
- 1 byte: Read Depth (DP), integer capped at 255
- 1 byte: Genotype (GT), encoded as:
  - 0 = Homozygous reference (0/0)
  - 1 = Heterozygous (0/1)
  - 2 = Missing or ambiguous
  - 3 = Homozygous alternate (1/1)

III. Result

A. Compression Efficiency

We evaluated the compression performance of our binary format using the chromosome 11 data from sample HG00157, sourced from the 1000 Genomes Project. The combined size of the original VCF and FASTA files was 573.85 MB. In contrast, the 2-bit encoded VCF file, which includes the full 2-bit encoded reference, 1-bit N mask, variant records, and metadata, required only 74.50 MB. This corresponds to a compression ratio of 87.02%.

The compression ratio was calculated as:

$$\text{Compression Rate} = \left(1 - \frac{74.50}{573.85}\right) \times 100 \approx 87.02\%$$

This result demonstrates that even with the inclusion of the full reference and N mask, our format achieves substantial reduction in file size compared to conventional formats.

B. Variant Inclusion Accuracy

To evaluate the accuracy of the 2-bit encoded format, we checked the number of variants that were correctly encoded into our encoded format. For chromosome 11 of sample HG00157, the original VCF file contained a total of 3,881,791

variants, including SNPs, insertions, and deletions. Among these, 3,870,547 variants were successfully encoded in the 2-bit encoded format.

Importantly, our format includes all variants regardless of genotype, including homozygous reference (0/0), heterozygous (0/1 or 1/0), homozygous alternate (1/1), and ambiguous or missing genotypes

- **Total variants in original VCF:** 3,881,791
- **Variants successfully encoded:** 3,870,547

These results demonstrate that our binary format provides accurate compression of variant data in the reference VCF file, making it suitable for sequence reconstruction and large-scale genomic data processing.

C. Genome Restoration Accuracy

To evaluate the accuracy of genome reconstruction from our binary format, we aligned the reconstructed FASTA sequence against the original reference genome using the Edlib alignment tool. The alignment was performed over a 10 Mbp region, and the resulting edit distance was 367,977, which means out of 10,000,000 bases, 367,977 were different. This corresponds to a sequence similarity of 96.32%. In other words, more than 96% of the reconstructed genome matches the reference at the base level.

$$\text{Similarity} = \left(1 - \frac{367,977}{10,000,000}\right) \times 100 \approx 96.32\%$$

The remaining 3.68% difference is primarily due to true biological variation—such as insertions and deletions—present in the individual genome but absent in the reference. Therefore, this deviation is expected and does not indicate an error in the reconstruction process.

This result confirms that our binary format enables highly accurate reconstruction of individual genomic sequences while preserving personal genetic variation.

However, as it only tests for 10,000,000 bases out of 134,741,342 bases, further testing is needed to fully validate the accuracy.

DISCUSSION

Our results demonstrate that the proposed binary format achieves a compression rate of approximately 87%, offering a substantial reduction in storage requirements compared to the traditional VCF + FASTA approach. This level of compression is particularly beneficial for large-scale sequencing or simulation-based genomic studies by improving disk space efficiency.

As shown in the reconstruction and alignment results, our format enables the restoration of complete personal genome sequences without relying on an external FASTA file. This self-contained design makes the format portable across different environments, enhances the reproducibility of genomic data, and useful for data analysis.

Compared to existing methodologies, one of the key distinctions between our format and existing solutions such as CRAM [3], [4] lies in reference dependency. While CRAM requires access to the reference FASTA during decoding or encoding, our format stores the entire reference within, allowing standalone restoration and improved portability.

Despite these strengths, several directions for future work remain. First, ensuring compatibility with existing libraries that rely on the standard VCF + FASTA model could broaden the applicability of our format in existing toolchains. Second, it would be valuable to evaluate whether this format can be extended to support alternative reference standards such as the Telomere-to-Telomere (T2T) assembly. Supporting longer insertions, structural variants, or multi-sample encoding could further increase the flexibility and adoption of this method.

These extensions would help make the format more practical and useful in real-world applications, and we plan to explore such improvements in future work.

REFERENCES

[1] Y. Turakhia et al., "Ultrafast sample placement on existing trees (UShER) empowers real-time phylogenetics for the SARS-CoV-2 pandemic," *Nature Genetics*, vol. 53, pp. 809–816, 2021.

[2] W. Liao et al., "A draft human pangenome reference," *Nature*, vol. 617, pp. 312–324, 2023.

[3] M. H. Fritz et al., "Efficient storage of high throughput DNA sequencing data using reference-based compression," *Genome Research*, vol. 21, no. 5, pp. 734–740, 2011.

[4] M. H. Fritz et al., "CRAM 3.1: advances in reference-based compression for high-throughput sequencing data," *bioRxiv*, 2021.