Unified Representation of Genetic Variants

Adrian Tan, Gonçalo R. Abecasis, and Hyun Min Kang

Supplementary Text

Lemma 1. For any variant, there exists a unique normalized VCF entry.

Proof : Let $(\boldsymbol{R} = R_1 R_2 \cdots R_n, \boldsymbol{A} = A_1 A_2 \cdots A_m)$ be the reference and alternate sequence of a bi-allelic variant $\boldsymbol{V}$, and let $(\boldsymbol{c}, b, \boldsymbol{r} = r_1 r_2 \cdots r_k, \boldsymbol{a} = a_1 a_2 \cdots a_l)$ be chromosome, base position, reference allele, and alternate allele of a VCF entry $\boldsymbol{E}$. Without loss of generality, assume that the reference genome consists of a single chromosome.

First, we show that at least one VCF entry $\boldsymbol{E}$ representing $\boldsymbol{V}$ exists. This can be trivially proven by setting $b = 1$, $\boldsymbol{r} = \boldsymbol{R}$, $\boldsymbol{a} = \boldsymbol{A}$

Second, we show that there is only one normalized VCF entry for a variant. If ***E** = (**c**, b, **r**, **a**)* represents ***V***, the following conditions must hold.

1) $r_1 = R_b, r_2 = R_{b+1}, \cdots, r_k = R_{b+k-1}$
2) $a_1 = A_b, a_2 = A_{b+1}, \cdots, a_l = A_{b+l-1}$
3) $m - n = l - k$
4) $R_1 = A_1, R_2 = A_2, \cdots, R_{b-1} = A_{b-1}$
5) $R_n = A_m, R_{n-1} = A_{m-1}, \cdots, R_{b+k} = A_{b+l}$

Suppose that there are two normalized VCF entries $\boldsymbol{E} = (\boldsymbol{c}, b, \boldsymbol{r}, \boldsymbol{a})$ and $\boldsymbol{E}' = (\boldsymbol{c}', b', \boldsymbol{r}', \boldsymbol{a}')$ that represents the same variant. Let k and k' be the length of ***r*** and ***r'***, respectively. Because both entries are parsimonious, by definition $k = k'$ must hold.

Because ***E*** and ***E'*** have the same allele length and are both normalized, by definition of left-alignment, *b = b'* must hold too.

Given the same ***R**, **A**, b,* and *k*, it is easy to show that the conditions (1)-(5) require that ***r*** and ***a*** are always uniquely determined. Therefore, ***E*** and ***E'*** must be identical VCF entries, and there is unique normalized VCF entry representing a variant.

The proof can be straightforwardly extended to multi-allelic variants or to reference genomes consisting of multiple chromosomes.

Lemma 2. A VCF entry is left aligned if and only if not all alleles end with the same nucleotide.

Proof : First, we need to show that, if $\boldsymbol{E}$ is left aligned, then $r_k \neq a_l$. Suppose that a VCF entry is left aligned (b is minimum when k is fixed), but $r_k = a_l$. Then we define

- $b' = b - 1$
- $\boldsymbol{r}' = R_{b-1} r_1 r_2 \cdots r_{k-1}$
- $\boldsymbol{a}' = A_{b-1} a_1 a_2 \cdots a_{l-1}$

Then $\boldsymbol{E}' = (\boldsymbol{c}, b', \boldsymbol{r}', \boldsymbol{a}')$ also represents $\boldsymbol{V}$ because $R_{b'+k} = r_k = a_l = A_{b'+l}$ and conditions (1)-(5) holds. Consequently, the entry is not left aligned ($b' < b$) contradicting the starting assumption. Therefore, if $\boldsymbol{E}$ is left aligned, $r_k \neq a_l$ holds.

Next, we want to show that $\boldsymbol{E}$ is always left aligned if $r_k \neq a_l$. Suppose that $\boldsymbol{E}$ is not left aligned. Then there exist a $\boldsymbol{E}' = (\boldsymbol{c}, b', \boldsymbol{r}', \boldsymbol{a}')$ such that $b' = b - i$ $(i > 0)$ and $\boldsymbol{r}'$ has length *k*. Then the conditions (1)-(5) above must hold. However, (5) cannot hold because

$$R_{b'+k+i-1} = R_{b+k-1} = r_k \neq a_l = A_{b+l-1} = A_{b'+l+i-1}$$

Thus, $\boldsymbol{E}$ is left aligned if and only if $r_k \neq a_l$.

Note that an exception to this Lemma can happen when $b = 1$ (i.e. not possible to shift the variant to the left), but we will ignore this case, which is extremely unlikely to happen for human autosomal and sex chromosomes, for the sake of simplicity.

Again, the proof can easily be generalized for multi-allelic variants and multi-chromosome reference genomes.

Lemma 3. A left-aligned VCF entry is parsimonious if and only if not all alleles start with the same nucleotide, or at least one allele has length 1.

Proof : First, suppose that one allele has length 1, then the entry is parsimonious because VCF files do not allow alleles with zero length.

Second, suppose that all allele lengths are greater than 1. If alleles of a left-aligned VCF entry start with different nucleotides, $r_1 \neq a_1$ holds. In addition, $r_k \neq a_l$ holds because it is left-aligned (lemma 2). Because both alleles end with different nucleotides, it is not possible to create a VCF entry $\boldsymbol{E}' = (\boldsymbol{c}, b', \boldsymbol{r}', \boldsymbol{a}')$ with smaller allele length without violating conditions (4) and (5). Consequently, the entry is parsimonious if it starts with different nucleotides.

Third, suppose that a left-aligned VCF entry $\boldsymbol{E}$ with minimum allele length greater than 1 is parsimonious but $r_1 = a_1$. Then we can define

- $b' = b + 1$
- $k' = k - 1$
- $\boldsymbol{r}' = r_2 \cdots r_k$
- $\boldsymbol{a}' = a_2 \cdots a_l$

Then $\boldsymbol{E}' = (\boldsymbol{c}, b', \boldsymbol{r}', \boldsymbol{a}')$ also represents $\boldsymbol{V}$ because $R_{b'-1} = r_1 = a_1 = A_{b'-1}$ and conditions (1)-(5) holds, and $\boldsymbol{E}$ is not parsimonious. Therefore, if $\boldsymbol{E}$ is parsimonious, $r_1 \neq a_1$ must hold.

To sum up the three parts together, a left-aligned VCF entry is parsimonious if and only if not all alleles start with the same nucleotide, or at least one allele has length 1.

Theorem 1. A VCF entry is normalized if and only if (1) not all alleles end with the same nucleotide, and (2) not all alleles start with the same nucleotide, or the allele length is 1.

Suppose that the two conditions hold. By (1) and Lemma 2, the entry is left-aligned. By (2) and Lemma 3, the left-aligned entry is parsimonious. Therefore, the VCF entry is both left-aligned and parsimonious, and thus normalized.

Suppose that a VCF entry is normalized. Because it is left-aligned, by Lemma 2, the alleles must not end with the same nucleotide. Because it is left-aligned and parsimonious, by Lemma 3, the alleles must not start with the same nucleotide unless the minimum allele length is 1. Therefore, a normalized VCF entry must satisfy both conditions.

Based on the proof for each direction, the Theorem 1 is true.