Genetic Variant Normalization

What we are normalizing

As input, we are getting records, describing genetic events (aka variants). There are several types of records in the input data:

- Point events, or SNV (Single Nucleotide Variations)
- Small sequence alterations: insertions, deletions (indels), etc.
- Large sequence alterations: copy number variations (CNV), structural variants (SV), etc.
- Chromosomal events (duplicate or missing chromosomes or large parts of thereof)

The sequence has to be normalized because events (variants) can be represented in ambiguous way. Ambiguity makes it impossible to match the events to the data associated with these events in knowledges sources. In general, only small sequence alteration events (the second bullet point above) can be ambiguous and thus, require normalization through disambiguation.

GA4GH Normalization Spec

See https://vrs.ga4gh.org/en/stable/impl-guide/normalization.html

Example of ambiguous records

For example take this deletion:

AT~~C~~CTA (i.e. a deletion of nucleotide C). This variant can be expressed in the following ways:

1. 2: TC>T
2. 3: CC>C
3. 2: TCC>TC

A more complex sequence alteration, like **AT~~T~~CT~~TG~~TGTA** has many more ways to be expressed. For normalization it would be best to split this variant into several in a way, that every smaller variant cannot be split further. However, even splitting is ambiguous and we have to follow a deterministic algorithm in the way we split.

Splitting SNV

In some cases, VCF file produced by some pipelines are merged into one. For example, we can find in VCF teh following sequence alteration variant:

```
12345678: AA>TT
```

For the purpose of annotation, we should split it into two SNV variants:

```
12345678: A>T
```

```
12345679: A>T
```

Multi-allelic variants

In VCFs, especially VCFs for joint calls (single VCF for multiple samples - the most common type for genetic analysis) some variants are multi-allelic. For example:

```
A>C,T
```

It can mean, that for one family member a reference nucleotide **A** is replaced with nucleotide **C,** while for another it is replaced with **T**. In rare cases it can mean that in a single sample, one copy of the DNA has **C** instead of **A**, and the second copy has **T**.

A more complex case is when a multi-allelic variant is not an SNV but a small sequence alteration event. For example:
`chr2:207467421-207467424 TACAC>T,TAC,TATACACACACAC`

corresponding to the following family genotypes (the data is public):

Ref	TACAC
Alt	T
Proband Genotype	T/T
Maternal Genotype	T/TAC
Paternal Genotype	T/TATACACACACAC

As each of the sub-variants in a multi-allelic variant can have different clinical consequences, they have to be annotated separately. The best practice is, for the purpose of annotation, to split all multi-allelic variants into a set of variants with the single alternative.