Variant Normalization Example

Consider this multi-sample VCF input record at chromosome 1 position 100. It lists four samples with their genotypes being; homozygous reference [AA /AA], heterozygous SNP [AA/AT], heterozygous insertion [AT/AAC] and heterozygous deletion [AA/A]:

```
#CHROM POS    REF    ALT       FORMAT  SAMPLE1       SAMPLE2        SAMPLE3        SAMPLE4
1      100    AA     AT,AAC,A  AT:AD   0/0:40,1,0,0  0/1:19,20,1,0  2/1:0:20,22,0  0/3:19,0,0,20
```

The OpenCB Variant Normalization process normalises first splits the record into three individual variants, one for each alternate allele;

```
#CHROM POS    REF    ALT
1      100    AA     AT
1      100    AA     AAC
1      100    AA     A
```

Each variant is then allele trimmed and positions updated;

```
#CHROM POS    REF    ALT
1      101    A      T
1      102    -      C
1      100    A      -
```

The final JSON representation of the Variant objects as stored in the OpenCGA database is as follows:

```
{
  "id" : "1:101:A:T",
  "chromosome" : "1",
  "start" : 101,
  "end" : 101,
  "reference" : "A",
  "alternate" : "T",
  "type" : "SNV",
  "studies" : [ {
    "files" : [ {
      "call" : {
        "variantId" : "1:100:AA:AT,AAC,A",
        "alleleIndex" : 0
      }
    } ],
    "secondaryAlternates" : [ {
      "chromosome" : "1",
      "start" : 102,
      "end" : 101,
      "reference" : "",
      "alternate" : "C",
      "type" : "INDEL"
    }, {
      "chromosome" : "1",
      "start" : 100,
      "end" : 100,
      "reference" : "A",
      "alternate" : "",
      "type" : "INDEL"
    } ],
    "sampleDataKeys" : [ "GT", "AD" ],
    "samples" : [ {
      "sampleId" : "SAMPLE1",
      "data" : [ "0/0", "40,1,0,0" ]
    }, {
      "sampleId" : "SAMPLE2",
      "data" : [ "0/1", "19,20,1,0" ]
    }, {
      "sampleId" : "SAMPLE3",
      "data" : [ "2/1", "0:20,22,0,0" ]
    }, {
      "sampleId" : "SAMPLE4",
      "data" : [ "0/3", "19,0,0,20" ]
    } ]
  } ]
}
{
  "id" : "1:102:-:C",
  "chromosome" : "1",
  "start" : 102,
  "end" : 101,
  "alternate" : "C",
```

```
  "type" : "INDEL",
  "studies" : [ {
    "files" : [ {
      "call" : {
        "variantId" : "1:100:AA:AT,AAC,A",
        "alleleIndex" : 1
      }
    } ],
    "secondaryAlternates" : [ {
      "chromosome" : "1",
      "start" : 101,
      "end" : 101,
      "reference" : "A",
      "alternate" : "T",
      "type" : "SNV"
    }, {
      "chromosome" : "1",
      "start" : 100,
      "end" : 100,
      "reference" : "A",
      "alternate" : "",
      "type" : "INDEL"
    } ],
    "sampleDataKeys" : [ "GT", "AD" ],
    "samples" : [ {
      "sampleId" : "SAMPLE1",
      "data" : [ "0/0", "40,0,1,0" ]
    }, {
      "sampleId" : "SAMPLE2",
      "data" : [ "0/2", "19,1,20,0" ]
    }, {
      "sampleId" : "SAMPLE3",
      "data" : [ "1/2", "0:20,0,22,0" ]
    }, {
      "sampleId" : "SAMPLE4",
      "data" : [ "0/3", "19,0,0,20" ]
    } ]
  } ]
}
{
  "id" : "1:100:A:-",
  "chromosome" : "1",
  "start" : 100,
  "end" : 101,
  "reference" : "A",
  "alternate" : "",
  "type" : "INDEL",
  "studies" : [ {
    "files" : [ {
      "call" : {
        "variantId" : "1:100:AA:AT,AAC,A",
        "alleleIndex" : 2
      }
    } ],
    "secondaryAlternates" : [ {
      "chromosome" : "1",
      "start" : 101,
      "end" : 101,
      "reference" : "A",
      "alternate" : "T",
      "type" : "SNV"
    }, {
      "chromosome" : "1",
      "start" : 102,
      "end" : 101,
      "reference" : "",
      "alternate" : "C",
      "type" : "INDEL"
    } ],
    "sampleDataKeys" : [ "GT", "AD" ],
    "samples" : [ {
      "sampleId" : "SAMPLE1",
      "data" : [ "0/0", "40,0,1,0" ]
    }, {
      "sampleId" : "SAMPLE2",
      "data" : [ "0/2", "19,0,20,1" ]
    }, {
      "sampleId" : "SAMPLE3",
      "data" : [ "3/2", "0:20,0,22,0" ]
    }, {
      "sampleId" : "SAMPLE4",
      "data" : [ "0/1", "19,20,0,0" ]
    } ]
  } ]
```

```
}
```