CATS
(Cancer Genomic Test Standardized) Format

Document for details

By Section of Genomic Data Management,
C-CAT

v1.1.0

2021/07/02

Contents

I. Introduction

I-1. Objectives

Currently, testing companies use different formats for data on gene alterations in comprehensive genomic profiling tests of cancer. The difference makes it difficult for a third party to annotate gene alterations with candidate drugs and clinical trials using the same software under a simple framework. To efficiently promote uniformity and homogeneity in the interpretation of cancer genomic test results, it is necessary to define a standardized format to present gene alteration data in cancer comprehensive genome profiling tests.

This document describes the CATS (cancer genomic test standardized) format, a standardized format for presenting gene alteration data in cancer comprehensive genomic profiling tests. The data schema of the CATS format is defined by the JSON definition file "schema.json", the specifications of which are explained in this document.

This format is used as follows: A laboratory that performs caner comprehensive genome profiling tests sends gene alteration data in the CATS format to a testing annotation organization such as C-CAT. The organization annotates gene alteration data in the CATS format with candidate drugs and clinical trials, taking into account the clinical data of patients. This format only applies to gene alteration data, not clinical data, because clinical data are stored in electronic medical records in hospitals and are unlikely to be accessible to testing companies.

I-2. Terms

- *Testing annotation organization*: An organization that receives gene alteration data in cancer comprehensive genome profiling tests from testing companies and clinical data from hospitals and uses the cancer knowledge base to annotate each gene alteration with candidate drugs and clinical trials. For example, C-CAT.
- *Testing annotation document*: A document that associates individual gene alterations in a cancer comprehensive genomic profiling test with candidate drugs and clinical trials for each patient. An example is the C-CAT Findings document. In contrast to a testing annotation document, a testing report document is a report of test results issued by a testing company.
- *Cancer knowledge base*: A database that associates gene alterations in cancer with candidate drugs and clinical trials. Examples include C-CAT CKDB (cancer knowledge database) and OncoKB (Chakravarty et al, 2017, JCO Precision Oncology).

I-3. About the condition field

- Required: Required field when the JSON parent tag exists.
- Optional: Recommended field that may be associated with drug and clinical trial information in testing annotation documents or that increases the accuracy of testing annotation documents according to testing reports issued by testing companies.

I-4. Format information

- Character code: UTF8
- Type: JSON
- Extension: json

II. Matters specific to C-CAT

II-1. Sending format of files

Gene alteration data in cancer comprehensive genomic profiling tests should be sent to C-CAT from the testing company in CATS (cancer genomic test standardized) format.

II-2. Scope of inputs

Quality-assured data on gene alterations (shortVariant, copyNumberAlteration, rearrangement, and otherBiomarker) are in the scope of inputs in CATS format. Please do not input alterations suspicious as false positives. You can choose whether or not to show alterations in C-CAT Findings by the tag ("reported") described below.

Please be sure to input data on gene alterations that are approved by the concerned authorities and be sure to output them in C-CAT Findings.

II-3. Requests

- It is recommended to provide inputs for as many optional fields as possible. As a result, more information on drugs or clinical trials may be added to C-CAT Findings, and C-CAT Findings based on laboratory testing reports will be more accurate. Additionally, more information linked to these fields may be added in future versions, even if such information does not appear in the current version of C-CAT Findings.
- Please try your best to input quality-assured data on all gene alterations (including those with the tag of "reported": false, as explained below) in CATS format. Otherwise, in case of any changes in the format or specifications of laboratory testing reports, C-CAT may send inquiries to the laboratory and the production of C-CAT Findings may be delayed.

II-4. Notes

Detailed precautions specific to C-CAT are indicated by '*' in the Description below.

III. metaData tag

This tag is used to define the metadata.

It contains 4 keys: schemaVersion, referenceGenome, configOptions, and comments.

Key	Condition	Data type	Description
metaData	required	object	Aggregation tag for metadata

III-1. schemaVersion key

Key	Condition	Data type	Description
schemaVersion	required	string regex: ^[0-9¥¥.]+$	Schema version of this format

III-2. referenceGenome tag

Key	Condition	Data type	Description
referenceGenome	required	object	For information on a reference genome sequence

III-2-1. Tags within referenceGenome tag

Key	Condition	Data type	Description
name	optional	string regex: ^.+$	Name of a reference genome sequence used in your test.
grcRelease	required	string regex: ^GRC.+$	GRC (Genome Reference Consortium) release ID of a reference genome sequence.
descriptions	optional	array (length: 0-N, string regex: ^.+$)	Description of the reference genome sequence in the name tag. See the contents tag within the comments tag for usable languages and new lines.

III-2-2. Examples of referenceGenome tag

(Example1. for NCBI)

```
"referenceGenome": {
  "name": "GRCh38.p13",
  "grcRelease": "GRCh38.p13",
  "descriptions": [
```

```
    "Homo sapiens (human) genome assembly GRCh37 (hg19) from the Genome Reference
Consortium."
  ]
}
```

(Example2. for UCSC)

```
"referenceGenome": {
  "name": "hg38Patch11",
  "grcRelease": "GRCh38.p11",
  "descriptions": [
    "GRCh38 Genome Reference Consortium Human Reference 38 (GCA_000001405.22))"
  ]
}
```

(Example3. for GDC)

```
"referenceGenome": {
  "name": "GRCh38.d1.vd1",
  "grcRelease": "GRCh38",
  "descriptions": [
    "Homo sapiens (human) genome assembly GRCh38 (hg38) from GDC, GRCh38.d1.vd1"
  ]
}
```

III-3. configOptions tag

This tag controls matching to cancer knowledge bases, such as C-CAT CKDB, and determines whether or not to hide values measured in your test from testing annotation documents such as the C-CAT Findings document.

Key	Condition	Data type	Description
configOptions	optional	object	Aggregation tag that controls matching to cancer knowledge bases and listing in testing annotation documents

III-3-1. Tags within configOptions tag

Key	Condition	Data type	Description
typeLabelsInterpretedAsKbAmplification	optional	array (length: 1-4, string) [choice]	The testing company's labels for gene alterations, which are interpreted as “amplification” (copy number amplification) in cancer knowledge bases. Choose from the following options (must not be duplicated in an array). ・"copyNumberAlterationType: amplification" ・"copyNumberAlterationType: gain"

			・"copyNumberAlterationType: duplication" ・"rearrangementType: duplication" (Default: "copyNumberAlterationType: amplification", "copyNumberAlterationType: gain", "copyNumberAlterationType: duplication")
typeLabelsInterpretedAsKbLoss	optional	array (length: 1-3, string) [choice]	The testing company's labels for gene alterations, which are interpreted as “loss” (copy number loss) in cancer knowledge bases. Choose from the following options (must not be duplicated in an array). ・"copyNumberAlterationType: loss" ・"copyNumberAlterationType: deletion" ・"copyNumberAlterationType: homozygous deletion" ・"rearrangementType: deletion" (Default: "copyNumberAlterationType: loss", "copyNumberAlterationType: deletion", "copyNumberAlterationType: homozygous deletion")
typeLabelsInterpretedAsKbGeneFusion	optional	array (length: 1-8, string) [choice]	The testing company's labels for gene alterations, which are interpreted as “geneFusion” (gene fusion) in cancer knowledge bases. Choose from the following options (must not be duplicated in an array). ・"rearrangementType: gene fusion" ・"rearrangementType: gene fusion and frameshift variant" ・"rearrangementType: bidirectional gene fusion" ・"rearrangementType: duplication" ・"rearrangementType: tandem duplication" ・"rearrangementType: deletion" ・"rearrangementType: inversion" ・"rearrangementType: truncation" ・"rearrangementType: other" (Default: "rearrangementType: gene fusion", "rearrangementType: gene fusion and frameshift variant", "rearrangementType: bidirectional gene fusion")

hideAlleleFrequency	optional	boolean	Variant allele frequency values will not be shown in testing annotation documents when this is true. * Please input false or do not include the key for information approved by the authorities. (Default: false)
hideCnaValue	optional	boolean	Same for the values of copy number alterations.
hideMsiValue	optional	boolean	Same for the values of micro-satellite instability (MSI).
hideTmbValue	optional	boolean	Same for the values of tumor mutation burden (TMB).
hideLohValue	optional	boolean	Same for the values of Loss of Heterozygosity (LOH).

III-3-2. Example of configOptions tag

(Example)

```
"configOptions": {
  "hideTmbValue": true,
  "hideLohValue": true,
  "typeLabelsInterpretedAsKbAmplification": [
    "copyNumberAlterationType: amplification",
    "copyNumberAlterationType: gain",
    "copyNumberAlterationType: duplication"
  ],
  "typeLabelsInterpretedAsKbLoss": [
    "copyNumberAlterationType: loss",
    "copyNumberAlterationType: deletion",
    "rearrangementType: deletion"
  ],
  "typeLabelsInterpretedAsKbGeneFusion": [
    "rearrangementType: gene fusion",
    "rearrangementType: gene fusion and frameshift variant",
    "rearrangementType: bidirectional gene fusion"
  ]
}
```

III-4. comments tag

You can comment on gene alterations (variants), biomarkers (otherBiomarkers), and information on sequencing samples (sequencingSamples). This tag contains the itemIds and contents tags.

Key	Condition	Data type	Description

comments	optional	array (length: 0-N, object)	Aggregation tag for comment information. Each object in the array must be unique.

III-4-1. Tags within comments tag

Key	Condition	Data type	Description
itemIds	required	array (length: 0-N, string regex: ^.+$)	The itemIds (multiple unique itemIds possible) to indicate alterations (variants), biomarkers (otherBiomarkers), and information on sequencing samples (sequencingSamples). Possible to comment on a test overall if you set the length of this key to be zero. * The content will not be shown in C-CAT Findings, if itemId is specified.
contents	required	array (length: 1-N, string regex: ^.+$)	The content of the comment for itemId. The description can be in English or Japanese. Please use array elements if you make new lines, because we ignore line feed codes in this tag.

III-4-2. Example of comments tag

(Example)

```
"comments": [
  {
    "itemIds": [],
```

Note: If the length of the itemIds array is zero, it represents a comment on this test overall.

```
    "contents": [
      "Amplification of the FGFR1 gene is observed in 5 to 20% of squamous cell carcinomas, and it has been reported that FGFR1 is sensitive to FGFR inhibitors in vitro.",
      "FGFR2 and FGFR3 gene activating mutations and FGFR3 gene fusions have been reported one after another, and their frequency is low at around 3%, but therapeutic effects with FGFR inhibitors are expected."
```

Note: If you want to make new lines in testing annotation documents, sentences or phrases should be separated as elements of an array.

```
    ]
  },
  {
    "itemIds": [
      "variant-1"
```

Note: The itemId of the mutation should be included to comment on a specific mutation.

```
    ],
    "contents": [
```

```
      "TSC1 functions independently of TSC2 and mTORC1."
    ]
  },
  {
    "itemIds": [
      "variant-1",
      "variant-5"
```
Note: You can comment on multiple mutations altogether by listing multiple itemIds.
```
    ],
    "contents": [
      "Although CD4 T cell percentage in Tsc1-/- mice was not strongly affected by Bim deficiency in vivo, TCR-mediated apoptosis of Tsc1-/- Bcl2l11-/- double knockout CD4 T cells was less pronounced compared with that of Tsc1-/- cells. (Kai Yang et al.)"
    ]
  }
]
```

IV. testInfo tag

This tag is used to provide test information.

Key	Condition	Data type	Description
testInfo	required	object	Aggregation tag for test information

IV-1. Tags within testInfo tag

Key	Condition	Data type	Description
testId	required	string regex: ^.+$	Any ID used by the testing company
testType	required	string [choice]	The combination of specimens used in the test. ・"tumor-only": test using tumor samples only ・"tumor and matched-normal": test using tumor and matched normal samples ・"tumor-only (cell-free)": test using cell-free tumor samples only ・"tumor (cell-free) and matched-normal": test using cell-free tumor samples and normal samples
softwareName	optional	string regex: ^.+$	Name of gene analysis software
softwareVersion	optional	string regex: ^.+$	Version of gene analysis software
panelName	required	string regex: ^.+$	Name of your genomic profiling (gene panel) test. * Please inform C-CAT beforehand if you want to use a genomic test NOT approved to use under National Health Insurance.
panelVersion	required	string regex: ^.+$	Version of your test

IV-2. Example of testInfo tag

(Example)

```
"testInfo": {
  "testId": "12345678901231900001",
  "testType": "tumor and matched-normal",
  "softwareName": "variant caller A",
```

```
  "softwareVersion": "ver.1.2",
  "panelName": "Multi-gene Panel A",
  "panelVersion": "ver.1.03-00"
}
```

V. variants tag

This tag is used to provide information on detected gene alterations. It contains shortVariants tag, copyNumberAlterations tag, and rearrangements tag.

Key	Condition	Data type	Description
variants	optional	object	Aggregation tag for information on alterations

V-1. shortVariants tag

This tag is used to define information on SNV (single nucleotide variation), insertion, deletion, delins (simultaneous insertion and deletion), indel (insertion and deletion), and MNV (multiple-nucleotide variation).

Key	Condition	Data type	Description
shortVariants	optional	array (length: 1-N, object)	Aggregation tag for information on single nucleotide variation (SNV), insertion (insertion), deletion (deletion), deletion and insertion (delins), insertion and deletion (indel), and multiple-nucleotide variant (MNV) of nucleotides. Each object in the array must be unique.

V-1-1. Tags within shortVariants tag

key	Condition	Data type	Description
itemId	required	string regex: ^.+$	An ID assigned to an alteration. It must be a unique string of characters within a single case.
chromosome	required	string regex: ^[a-zA-Z0-9_¥¥-]+$	Chromosome number
position	required	integer	Physical position in a chromosome. Please use the 1-based coordinate system and describe according to VCF v4.3 (For example, as stated on page 13 of VCF v4.3, when the reference base is atCga and the variant base is at-ga, and C position in

			the reference base is 3, denote the physical position as "position": 2, "referenceAllele": "TC", and "alternateAllele": "T". As described below, "referenceAllele" represents a reference base and "alternateAllele" represents a variant base.)
referenceAllele	required	string regex: ^[ACGTN]+$	Reference base. Please describe according to VCF v4.3 (see the example above).
alternateAllele	required	string regex: ^[ACGTN¥¥*]+$	Variant base. Please describe according to VCF v4.3 (see the example above). Multi-alleles – tri-alleles and more – should be listed as different elements in the shortVariants tag. In that case, indicate the itemIds and that the variants are multi-alleles in the comments tag.
alternateAlleleFrequency	required	number	Variant allele frequency (ranging from 0 to 1)
totalReadDepth	optional	integer	Total read depth (minimum value 1)
alternateAlleleReadDepth	optional	integer	Variant allele read depth (minimum value 1)
variantType	optional	string [choice]	Short variant type written in the report by the testing company. Select one from: ・"SNV" ・"insertion" ・"deletion" ・"delins" ・"indel" ・"MNV"

			* Please inform C-CAT beforehand if you use other types.
transcripts	required	array (length: 1-N, object)	Information on the representative transcript. Each object in the array must be unique. (One gene should have one representative transcript.)
transcriptId	required	string regex: ^[^¥¥s]+$	Transcription product ID (e.g., NM_000368.4). It is strongly recommended to include sub-numbers (underlined in red above) for accurate computation. If no transcriptId is present due to a mutation in an intergenic region, then you can define "transcriptId": null.
transcriptDatabaseName	required	string [choice]	Database of the transcript ID. Select from the following two options. ・"RefSeq" ・"Ensembl" If the transcriptId is null, then you can define "transcriptDatabaseName": null.
transcriptDatabaseVersion	optional	string regex: ^. +$	Version of the database. * If this is not defined, the version determined by C-CAT will be used.
geneSymbol	required	string regex: ^[^¥¥s]+$	Name of the gene (gene symbol), as written in the testing report document by the testing company. Enter "geneSymbol": null if a gene does not exist. Enter an associated gene such as “TERT promoter,” if any.

strand	optional	string [choice]	Direction of transcription. If the orientation is the same as that of the reference genome sequence, it is "+"; if the orientation is opposite, it is "-". If the transcriptId is null, then you can define "strand": null.
cdsChange	required	string regex: ^. +$	Enter changes at the DNA level, as written in the testing company's report. Notation in HGVS is recommended. When an RNA is not transcribed as in an intergenic region, you can input null (in addition to the notation of non-coding regions such as n.*).
aminoAcidsChange	required	string regex: ^. +$	Enter changes at the protein level, as written in the testing company's report. Notation in HGVS is recommended. If no change is observed in amino acids as in an untranslated region, you can input null.
calculatedEffects	optional	array (length: 0-N, string regex: ^. +$)	Note the effects of alterations on transcripts, such as "splicing_variant", using Sequence Ontology terms. This field corresponds to "Effect (Sequence Ontology)" of the snpEff tool and VCF "Func.refGene" of the annovar tool. For annovar, this is explained at Output file 1 (refSeq gene annotation) on Gene-based Annotation in User Guide, whereby terms provided

			by annovar can be converted to Sequence Ontology terms. One term should be assigned to each element of an array. Each string in the array must be unique.
testMethod	required	string [choice]	Derived from ・"DNA-seq": DNA sample ・"RNA-seq": RNA sample
variantOrigin	optional	string [choice]	Somatic or germline origin. * C-CAT uses a different cancer knowledge base, according to whether alterations are somatic or germline. If no input is provided, the knowledge base for somatic alterations is used. ・"somatic": derived from somatic cells ・"germline": derived from germline cells ・"likely somatic": typically in a tumor-only test, the alteration is likely to be somatic, and the knowledge base for somatic alterations is used. ・"likely germline": typically in a tumor-only test, the alteration is likely to be germline, and the knowledge base for germline alterations is used.
reported	required	boolean	Whether or not the alteration is reported in testing company’s report or similar documents. When it is true, annotation is made based on the cancer knowledge base.

V-1-2. Example of shortVariants tag

(Example1. for SNV)

```
{
  "itemId": "variant-1",
```

Note: The itemId is a character string at the discretion of the testing company for the detected variant.

```
  "chromosome": "9",
  "position": 135781005,
  "referenceAllele": "C",
  "alternateAllele": "G",
```

Note: Describe "position" and "referenceAllele", "alternateAllele" according to the rules of VCF v4.3.

```
  "alternateAlleleFrequency": 0.54,
  "alternateAlleleReadDepth": 108,
  "totalReadDepth": 200,
  "variantType": "SNV",
  "transcripts": [
    {
      "transcriptId": "NM_000368.4",
      "transcriptDatabaseName": "RefSeq",
      "transcriptDatabaseVersion": "Release 99",
      "geneSymbol": "TSC1",
      "cdsChange": "c.1960C>G",
      "aminoAcidsChange": "p.Q654E",
      "calculatedEffects": [
        "missense_variant"
      ]
    }
  ],
  "testMethod": "DNA-seq",
  "variantOrigin": "somatic",
  "reported": true
}
```

(Example2. for insertion)

```
{
  "itemId": "variant-4",
  "chromosome": "8",
  "position": 37553560,
  "referenceAllele": "A",
  "alternateAllele": "AAGCGGC",
  "alternateAlleleFrequency": 0.4953,
  "alternateAlleleReadDepth": 368,
  "totalReadDepth": 743,
  "variantType": "insertion",
  "transcripts": [
    {
      "transcriptId": "NM_025069.2",
      "transcriptDatabaseName": "RefSeq",
```

```
      "transcriptDatabaseVersion": "Release 99",
      "geneSymbol": "ZNF703",
      "cdsChange": "c.63_64insAGCGGC",
      "aminoAcidsChange": "G21_G22insSG"
    }
  ],
  "testMethod": "DNA-seq",
  "variantOrigin": "somatic",
  "reported": true
}
```


(Example3. for deletion)


```
{
  "itemId": "variant-2",
  "chromosome": "1",
  "position": 27097751,
  "referenceAllele": "TC",
  "alternateAllele": "T",
  "alternateAlleleFrequency": 0.12,
  "alternateAlleleReadDepth": 32,
  "totalReadDepth": 266,
  "variantType": "deletion",
  "transcripts": {
    "transcriptId": "ENST00000324856.13",
    "transcriptDatabaseName": "Ensembl",
    "transcriptDatabaseVersion": "v99",
    "geneSymbol": "ARID1A",
    "cdsChange": "c.3340delC",
    "aminoAcidsChange": "p.P1115fs*46",
    "calculatedEffects": [
      "frameshift_variant"
    ]
  },
  "testMethod": "DNA-seq",
  "variantOrigin": "somatic",
  "reported": true
}
```


(Example4. for delins)


```
{
  "itemId": "variant-6",
  "chromosome": "1",
  "position": 26696982,
  "referenceAllele": "GC",
  "alternateAllele": "TT",
  "alternateAlleleFrequency": 0.25,
  "alternateAlleleReadDepth": 52,
  "totalReadDepth": 524,
  "variantType": "delins",
```

```
  "transcripts": {
    "transcriptId": "NM_007294.4",
    "transcriptDatabaseName": "RefSeq",
    "transcriptDatabaseVersion": "Release 99",
    "geneSymbol": "BRCA1",
    "cdsChange": "c.579_580delinsTT",
    "aminoAcidsChange": "p.E193_P194delinsDS"
  },
  "testMethod": "DNA-seq",
  "variantOrigin": "somatic",
  "reported": true
}
```

(Example5. for "TERT promoter")

```
{
  "itemId": "variant-5",
  "chromosome": "5",
  "position": 1295113,
  "referenceAllele": "G",
  "alternateAllele": "A",
  "alternateAlleleFrequency": 0.163,
  "alternateAlleleReadDepth": 15.9,
  "totalReadDepth": 92,
  "variantType": "SNV",
  "transcripts": {
    "transcriptId": "ENST00000310581.9",
    "transcriptDatabaseName": "Ensembl",
    "transcriptDatabaseVersion": "Release 99",
    "geneSymbol": "TERT",
    "cdsChange": "n.1295113C>T",
    "aminoAcidsChange": null,
    "calculatedEffects": [
        "TF_binding_site_variant"
    ]
  },
  "testMethod": "DNA-seq",
  "variantOrigin": "somatic",
  "reported": true
}
```

V-2. copyNumberAlterations tag

This tag is used to provide information on copy number alterations (CNAs).

Key	Condition	Data type	Description

copyNumberAlterations	optional	array (length: 1-N, object)	Aggregation tag for information on copy number alterations (CNAs). Each object in the array must be unique.

V-2-1. Tags within copyNumberAlterations tag

Key	Condition	Data type	Description
itemId	required	string regex: ^.+$	An ID assigned to an alteration. It must be a unique string of characters within a single case.
chromosome	optional	string regex: ^[a-zA-Z0-9_¥¥-]+$	Chromosome number
startPosition	optional	integer	Physical starting position in a chromosome. Please use the 1-based coordinate system.
endPosition	optional	integer	Physical ending position in a chromosome. Please use the 1-based coordinate system.
copyNumberMetrics	optional	array (length: 0-N, object)	Measurements and units of the copy number alteration. Array of objects composed of two keys, value and unit. If there are two or more values in different units, register them as an array with the length of 2 or more. Each object in the array must be unique.
value	required	number	Aberrated copy number measurement
unit	required	string [choice]	Unit for measured value. Select one from: ・"absolute copy number": Absolute copy number ・"fold-change": Ratio of (standardized) reading depth of tumor samples to normal samples

			・"log2 fold-change": log2 transformation of "fold-change" ・"fraction-of-gene": fraction of a CNA region to the gene region of interest * Please consult C-CAT if you want to use other units.
copyNumberAlterationType	required	string [choice]	CNA type written in the report by the testing company. Select one from: ・"amplification" ・"gain" ・"duplication" ・"loss" ・"deletion" ・"homozygous deletion" ・"neutral" * Please consult C-CAT if you want to use other types.
transcripts	required	array (length: 1-N, object)	Refer to the description in the shortVariants tag.
transcriptId	optional	string regex: ^[^¥¥s]+$	Refer to the description in the shortVariants tag.
transcriptDatabaseName	optional	string [choice]	Refer to the description in the shortVariants tag. If the transcriptId is entered, this key is recommended to input as well.
transcriptDatabaseVersion	optional	string regex: ^. +$	Refer to the description in the shortVariants tag.
geneSymbol	required	string regex: ^[^¥¥s]+$	Refer to the description in the shortVariants tag.
strand	optional	string [choice]	Refer to the description in the shortVariants tag.
cdsChange	optional	string regex: ^.+$	Refer to the description in the shortVariants tag.

aminoAcidsChange	optional	string regex: ^.+$	Refer to the description in the shortVariants tag.
calculatedEffects	optional	array (length: 0-N, string regex: ^.+$)	Refer to the description in the shortVariants tag.
testMethod	required	string [choice]	Refer to the description in the shortVariants tag.
variantOrigin	optional	string [choice]	Refer to the description in the shortVariants tag.
reported	required	boolean	Refer to the description in the shortVariants tag.

V-2-2. Example of copyNumberAlterations tag

(Example)

```
{
  "itemId": "variant-9",
  "chromosome": "1",
  "startPosition": 8921059,
  "endPosition": 8939151,
  "copyNumberMetrics": [
    {
      "value": 0.2309,
      "unit": "fold-change"
    },
    {
      "value": -2.1147,
      "unit": "log2 fold-change"
    }
  ],
  "copyNumberAlterationType": "loss",
  "transcripts": [
    {
      "transcriptDatabaseName": "RefSeq",
      "transcriptDatabaseVersion": "Release 99",
      "geneSymbol": "ENO1"
    }
  ],
  "testMethod": "DNA-seq",
  "variantOrigin": "somatic",
  "reported": false
}
```

V-3. rearrangements tag

This tag is used to provide information on rearrangements such as fusions, duplications, large deletions, and inversions.

Key	Condition	Data type	Description
rearrangements	optional	array (length: 1-N, object)	Aggregation tag for information on rearrangements, such as fusions, duplications, large deletions, and inversions. Each object in the array must be unique.

V-3-1. Tags within rearrangements tag

Key	Condition	Data type	Description
itemId	required	string regex: ^.+$	An ID assigned to an alteration. It must be a unique string of characters within a single case.
breakends	required	array (length: 2, object)	Two breakends of the rearrangement. Each object in the array must be unique.
chromosome	required	string regex: ^[a-zA-Z0-9_¥¥-]+$	Chromosome number
startPosition	required	integer	Physical starting position in a chromosome. Please use the 1-based coordinate system.
endPosition	required	integer	Physical ending position in a chromosome. Please use the 1-based coordinate system.
matePieceLocation	optional	string [choice]	If the sequence of interest is bound to another sequence on the upstream (or downstream) of this sequence of interest along the reference genome sequence, it is input as "upstream" (or “downstream”).

			Regarding gene fusions and other rearrangements ("rearrangementType": "other"), it is strongly recommended to input this key for accurately identifying the genomic changes. See "VII-2. matePieceLocation" below for a detailed explanation.
totalReadCount	optional	integer	Number of the total reads of a breakend.
alternateAlleleFrequency	optional	number	Variant allele frequency of a breakend (ranging from 0 to 1). Please input a variant allele frequency calculated for each breakend.
transcripts	Required	array (length: 1-N, object)	Refer to the description in the shortVariants tag.
transcriptId	optional	string regex: ^[^¥¥s]+$	Refer to the description in the shortVariants tag.
transcriptDatabaseName	optional	string [choice]	Refer to the description in the shortVariants tag. If the transcriptId is entered, this key is recommended to input as well.
transcriptDatabaseVersion	optional	string regex: ^.+$	Refer to the description in the shortVariants tag.
geneSymbol	required	string regex: ^[^¥¥s]+$	Refer to the description in the shortVariants tag.
strand	optional	string [choice]	Refer to the description in the shortVariants tag.
cdsChange	optional	string regex: ^.+$	Refer to the description in the shortVariants tag.

aminoAcidsChange	optional	string regex: ^.+$	Refer to the description in the shortVariants tag.
calculatedEffects	optional	array (length: 0-N, string regex: ^.+$)	Refer to the description in the shortVariants tag.
genePairs	optional	array (length: 0-N, string regex: ^.+$)	A transcriptionally ordered pair of gene names that flank the break-ends in the gene rearrangement. Connect the geneSymbols listed in transcripts with a "-" in the transcriptional direction. For example, for geneSymbols A and B, if A is upstream of transcription and B is downstream of transcription, it is represented as "A-B". * This directional information is taken into account in the cancer knowledge base search. In the absence of this tag, the pair of geneSymbols is treated as if there is no information on the transcriptional direction and the knowledge base is searched for the unordered pairs of gene names. * If this tag is omitted (and there is no mention in matePieceLocation), any pair of geneSymbols will be searched for unordered pairs of geneSymbols, but if you want to control the search in detail, please describe it as ["A-B", "B-A", "C-B"]. Each string in the array must be unique.
insertedSequence	optional	string regex: ^[ACGTN]+$	Sequence inserted between the two breakends of the genome sequence. If an inserted sequence does not exist, input null.
supportingReadCount	optional	integer	Number of support reads

totalReadCount	optional	integer	Number of total reads
alternateAlleleFrequency	optional	number	Variant allele frequency of a rearrangement (ranging from 0 to 1). Please input a value calculated for a rearrangement.
expressionLevelMetrics	optional	array (length: 0-N, object)	Information on expression levels in RNA-seq. Each object in the array must be unique.
value	required	number	Value of an expression level
unit	required	string [choice]	Unit of an expression level. Select one from the following options: ・"TPM" ・"FPKM" ・"FPM" ・"RPKM" ・"RPM" * Please consult C-CAT if you want to use other units.
rearrangementNames	optional	array (length: 0-N, string regex: ^. +$)	Individual rearrangement name given by the testing company. For example, "EML4-ALK fusion" Each string in the array must be unique.
rearrangementType	required	string [choice]	Rearrangement type written in the report by the testing company. Select one from: ・"gene fusion" ・"gene fusion and frameshift variant" ・"bidirectional gene fusion" ・"duplication" ・"deletion" ・"inversion" ・"truncation" ・"splice variant" ・"tandem duplication" ・"other" * Please consult C-CAT if you want to use other types.

testMethod	required	string [choice]	Refer to the description in the shortVariants tag.
variantOrigin	optional	string [choice]	Refer to the description in the shortVariants tag.
reported	required	boolean	Refer to the description in the shortVariants tag.

V-3-2. Example of rearrangements tag

(example1. In cases where genePairs is present)


```
{
  "itemId": "variant-13",
  "breakends": [
    {
      "chromosome": "2",
      "startPosition": 42510050,
      "endPosition": 42510050,
      "matePieceLocation": "downstream",
      "transcripts": [
        {
          "transcriptDatabaseName": "RefSeq",
          "transcriptDatabaseVersion": "Release 99",
          "geneSymbol": "EML4"
        }
      ]
    },
    {
      "chromosome": "2",
      "startPosition": 29445240,
      "endPosition":  29445240,
      "matePieceLocation": "upstream",
      "transcripts": [
        {
          "transcriptDatabaseName": "RefSeq",
          "transcriptDatabaseVersion": "Release 99",
          "geneSymbol": "ALK"
        }
      ]
    }
  ],
  "genePairs": [
    "EML4-ALK"
  ],
  "supportingReadCount": 30,
  "totalReadCount": 430,
  "alternateAlleleFrequency": 0.07,
  "rearrangementType": "other",
  "variantOrigin": "somatic",
```

```
  "testMethod": "DNA-seq",
  "reported": false
}
```

(example2. When describing the totalReadCount and alternateAlleleFrequency for each breakend)

```
{
  "itemId": "variant-13",
  "breakends": [
    {
      "chromosome": "2",
      "startPosition": 42510050,
      "endPosition": 42510050,
      "matePieceLocation": "downstream",
      "totalReadCount": 330,
      "alternateAlleleFrequency": 0.06,
      "transcripts": [
        {
          "transcriptDatabaseName": "RefSeq",
          "transcriptDatabaseVersion": "Release 99",
          "geneSymbol": "EML4"
        }
      ]
    },
    {
      "chromosome": "2",
      "startPosition": 29445240,
      "endPosition":  29445240,
      "matePieceLocation": "upstream",
      "totalReadCount": 570,
      "alternateAlleleFrequency":0.07,
      "transcripts": [
        {
          "transcriptDatabaseName": "RefSeq",
          "transcriptDatabaseVersion": "Release 99",
          "geneSymbol": "ALK"
        }
      ]
    }
  ],
  "genePairs": [
    "EML4-ALK"
  ],
  "supportingReadCount": 30,
  "rearrangementType": "other",
  "variantOrigin": "somatic",
  "testMethod": "DNA-seq",
  "reported": false
}
```

(example3. In cases where insertedSequence is present)


```
{
  "itemId": "variant-14",
  "breakends": [
    {
      "chromosome": "14",
      "startPosition": 234567,
      "endPosition": 234567,
      "matePieceLocation": "downstream",
      "transcripts": [
        {
          "transcriptDatabaseName": "RefSeq",
          "transcriptDatabaseVersion": "Release 99",
          "geneSymbol": null
        }
      ]
    },
    {
      "chromosome": "2",
      "startPosition": 321672,
      "endPosition":  321672,
      "matePieceLocation": "upstream",
      "transcripts": [
        {
          "transcriptDatabaseName": "RefSeq",
          "transcriptDatabaseVersion": "Release 99",
          "geneSymbol": "LINC01865"
        }
      ]
    }
  ],
  "insertedSequence": "GTNNNNNCAT",
  "supportingReadCount": 30,
  "alternateAlleleFrequency": 0.07,
  "rearrangementType": "other",
  "variantOrigin": "somatic",
  "testMethod": "DNA-seq",
  "reported": false
}
```

VI. otherBiomarkers tag

This tag is used to provide information on biomarkers other than alterations defined in the variants tag. Currently, Micro-Satellite Instability (MSI), Tumor Mutation Burden (TMB), and Loss Of Heterozygosity (LOH) can be noted.

Key	Condition	Data type	Description
otherBiomarkers	optional	array (length: 0-N, object)	Aggregation tag for information on biomarkers. Each object in the array must be unique.

VI-1. Tags within otherBiomarkers tag

Key	Condition	Data type	Description
itemId	required	string regex: ^. +$	An ID assigned to a biomarker. It must be a unique string of characters within a single case.
biomarkerType	required	string [choice]	Type of biomarkers. Select one from: ・"TMB": Tumor Mutation Burden ・"MSI": Micro-Satellite Instability ・"LOH": Loss Of Heterozygosity * Inform C-CAT for other biomarkers.
biomarkerMetrics	optional	array (length: 0-N, object)	Inspection values and units. Each object in the array must be unique.
value	required	number	Measured value e.g.) 5.15
unit	required	string regex: ^. +$	Unit for the measured value. Units may vary depending on test types. e.g.) %
state	optional	string [choice]	Select the biomarker status from the following options. ・"high" ・"low" ・"intermediate" ・"cannot be determined" ・"stable" If the test was performed but the result is not listed above, describe null.

			* Inform C-CAT for other options.
description	optional	array (length: 0-N, string regex: ^.+$)	Descriptions such as how to obtain the testing value and the meaning of the value. Refer to the contents tag within the comments tag for usable languages and new lines.
biomarkerOrigin	optional	string [choice]	Refer to the description in the shortVariants tag.
reported	required	boolean	Refer to the description in the shortVariants tag.

VI-2. Example of otherBiomarkers tag

(Example)

```
"otherBiomarkers": [
  {
    "itemId": "biomarker-1",
    "biomarkerType": "MSI",
    "biomarkerMetrics": [
      {
        "value": 5.15,
        "unit": "%"
      },
      {
        "value": 2,
        "unit": "MSIsensor score"
      }
    ],
```

Note: If one inspection item has values in multiple units, use array notation in the "biomarkerMetrics" tag.

```
    "state": "stable",
    "descriptions": [
      "MSI sensor score 10 points or more was MSI-H, 3 points or more and less than 10 points was indeterminate (MSI-I), and less than 3 points was microsatellite stable (MSS).",
      "https://www.gi-cancer.net/gi/ronbun/archives/201901-01.html"
    ],
    "reported": true
  },
  {
    "itemId": "biomarker-2",
```

```
    "biomarkerType": "TMB",
    "biomarkerMetrics": [
      {
        "value": 34.5680122,
        "unit": "Muts/Mb"
      }
    ],
    "state": "high",
    "reported": true
  },
  {
    "itemId": "biomarker-3",
    "biomarkerType": "LOH",
    "biomarkerMetrics": [
      {
        "value": 24.14,
        "unit": "%"
      }
    ],
    "state": "neutral",
    "reported": true
  }
]
```

VII. compositeBiomarkers tag

This tag provides information on composite markers (e.g., combination of gene alterations; fusion composed of three genes) that are represented by the combinations of elements in the shortVariants, copyNumberAlterations, and rearrangements tags.

Key	Condition	Data type	Description
compositeBiomarkers	optional	array (length: 0-N, object)	Aggregation tag for information on composite markers. Each object in the array must be unique.

VII-1. Tags within compositeBiomarkers tag

Key	Condition	Data type	Description
itemId	required	string regex: ^. +$	An ID assigned to a composite marker. It must be a unique string of characters within a single case.
componentItemIds	required	array (length: 2-N, string regex: ^. +$)	Array of component gene alterations (itemIds). Each string in the array must be unique.
biomarkerNames	required	array (length: 1-N, string regex: ^. +$)	The name of the composite marker, as listed in the testing company's report. Each string in the array must be unique.
descriptions	optional	array (length: 0-N, string regex: ^. +$)	Description on composite markers Refer to the contents tag within the comments tag for usable languages and new lines.
reported	required	boolean	Refer to the description in the shortVariants tag.

VII-2. Example of compositeBiomarkers tag

(Example)

```
"compositeBiomarkers": [
  {
    "itemId": "composite-1",
    "componentItemIds": [
        "variant-14",
        "variant-15"
      ],
    "biomarkerNames": [
        "BRAF-NRG1-ALK fusion"
    ],
    "descriptions": [
      "Three genes are fused together to produce the fusion gene BRAF-NRG1-ALK."
    ],
    "reported": true
  },
]
```

VIII. sequencingSamples tag

This tag is used to provide the information on sequencing samples results in the NGS run.

Key	Condition	Data type	Description
sequencingSamples	optional	array (length: 1-N, object)	Aggregation tag for information on sequencing samples results. Each object in the array must be unique.

The maximum length of this array is 4, since the tumorOrNormal tag and the testMethod tag can each take two values “tumor” or “normal” and “dnaSeq” or “rnaSeq”, respectively.

VIII-1. Tags within sequencingSamples tag

Key	Condition	Data type	Description
itemId	required	string regex: ^. +$	The ID of the item. It must be a unique string of characters within a single case.
tumorOrNormal	required	string [choice]	Whether the sequencing samples results are of tumor specimen or normal specimen. ・"tumor" ・"normal"
testMethod	required	string [choice]	Whether the sequencing samples results are derived from DNA or RNA samples. ・"DNA-seq": DNA sample ・"RNA-seq": RNA sample
duplicateReadsPercentage	optional	number	Percentage of duplicate reads in total reads
mappedReadsPercentage	optional	number	Percentage of mapped reads in total reads
meanReadDepth	optional	number	Mean read depth
medianReadDepth	optional	number	Median read depth
suspectedSampleStates	optional	array (length: 0-N, string) [choice]	The suspicious state of a DNA or RNA sample, such as "contaminated" (must not be duplicated in an array). ・"contaminated": Possibility of contamination ・"deaminated": Possibility of notable cytosine deamination in FFPE DNA

			・"fragmented": Possibility of notable (FFPE DNA) fragmentation. ・"degraded": Possibility of notable (RNA) degradation

VIII-2. Example of sequencingSamples tag

(Example)

```
"sequencingSamples": [
  {
    "itemId": "sequence-1",
    "tumorOrNormal": "tumor",
    "testMethod": "DNA-seq",
    "duplicateReadsPercentage": 91.52,
    "mappedReadsPercentage": 87.31,
    "meanReadDepth": 247.8,
    "medianReadDepth": 238
  },
  {
    "itemId": "sequence-2",
    "tumorOrNormal": "tumor",
    "testMethod": "RNA-seq",
    "suspectedSampleStates": [
      "degraded"
    ]
  },
  {
    "itemId": "sequence-3",
    "tumorOrNormal": "normal",
    "testMethod": "DNA-seq"
  }
]
```

IX. Other notes

Precautions.

IX-1. itemId

The value of itemId must be unique within the file. The value can be any string.

IX-1-1. Example of itemId description

Examples of itemId values for each tag are as follows.

- For variants tag

```
"itemId": "variant-1"
```

```
"itemId": "variant-2"
```

```
"itemId": "variant-3"
```

- For otherBiomarkers tag

```
"itemId": "biomarker-1"
```

```
"itemId": "biomarker-2"
```

```
"itemId": "biomarker-3"
```

- For sequencingSamples tag

```
"itemId": "sequence-1"
```

```
"itemId": "sequence-2"
```

```
"itemId": "sequence-3"
```

- For compositeBiomarkers tag

```
"itemId": "composite-1"
```

```
"itemId": "composite-2"
```

```
"itemId": "composite-3"
```

IX-2. matePieceLocation

Here, we provide an explanation for matePieceLocation in the breakends tag. We assume that “upstream” and “downstream” correspond to the direction of decreasing and increasing positional coordinates along a chromosome in a reference genome sequence, respectively.

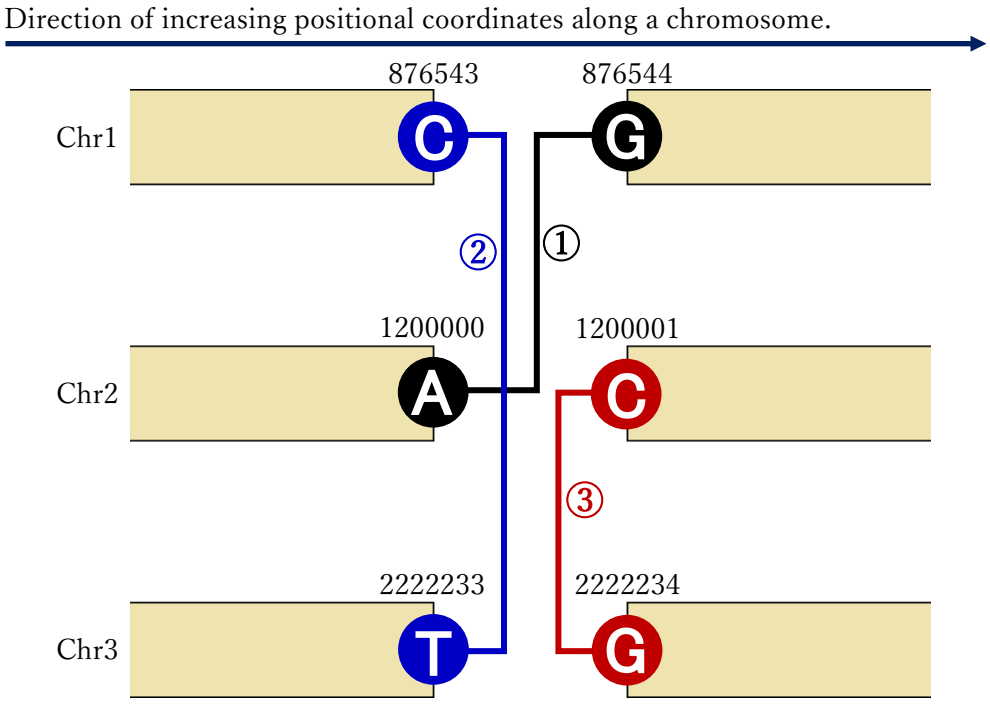
Note: These “upstream” and “downstream” are different from those in the transcriptional direction.

IX-2-1. Example of matePieceLocation description

We illustrate how to input matePieceLocation and present representations in VCF format (v4.3). The examples below are modified from the figure in the following documents:

The Variant Call Format Specification VCF v4.3 and BCF v2.2
https://samtools.github.io/hts-specs/VCFv4.3.pdf

Example of rearrangements:



Description in VCF format:

The representations in VCF v4.3 for the figure above are as follows: The numbers in the leftmost column in the table below correspond to the numbers in the figure above.

	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
①	1	876544	bnd_V	G	]2:1200000]G	6	PASS	SVTYPE=BND
	2	1200000	bnd_U	A	A[1:876544[	6	PASS	SVTYPE=BND

②	1	876543	bnd_W	C	C]3:2222233]	6	PASS	SVTYPE=BND
	3	2222233	bnd_Y	T	T]1:876543]	6	PASS	SVTYPE=BND
③	2	1200001	bnd_X	C	[3:2222234[C	6	PASS	SVTYPE=BND
	3	2222234	bnd_Z	G	[2:1200001[G	6	PASS	SVTYPE=BND

Descriptions in the standardized format:

・Example①

For the chromosome 2 junction point, the downstream sequence of the junction point is replaced with another sequence; therefore, the value of matePieceLocation is "downstream". In contrast, for the chromosome 1 junction point, the upstream sequence of the junction point is replaced with another sequence; therefore, the value of matePieceLocation is "upstream".

(Example①)

```
"breakends": [
  {
    "chromosome": "2",
    "startPosition": 1200000,
    "endPosition": 1200000,
    "matePieceLocation": "downstream"
  },
  {
    "chromosome": "1",
    "startPosition": 876654,
    "endPosition": 876654,
    "matePieceLocation": "upstream"
  }
]
```

・Example②

For the chromosome 1 junction point, the downstream sequence of the junction point is replaced with another sequence; therefore, the value of matePieceLocation is "downstream". The same is true for chromosome 3; thus, the value of matePieceLocation is "downstream".

(Example②)

```
"breakends": [
  {
    "chromosome": "1",
    "startPosition": 876543,
    "endPosition": 876543,
    "matePieceLocation": "downstream"
  },
  {
    "chromosome": "3",
```

```
    "startPosition": 2222233,
    "endPosition": 2222233,
    "matePieceLocation": "downstream"
  }
]
```

・Example③

For the chromosome 2 junction point, the upstream sequence of the junction point is replaced with another sequence; therefore, the value of matePieceLocation is "upstream". The same is true for the chromosome 3 junction point; thus, the value of matePieceLocation is "upstream".

```
(Example③)
"breakends": [
  {
    "chromosome": "2",
    "startPosition": 1200001,
    "endPosition": 1200001,
    "matePieceLocation": "upstream"
  },
  {
    "chromosome": "3",
    "startPosition": 2222234,
    "endPosition": 2222234,
    "matePieceLocation": "upstream"
  }
]
```

X. For inquires

Please contact the C-CAT Help Desk.

E-Mail: helpdesk_c-cat@ml.res.ncc.go.jp