BioCompute Object for Regulatory Review

BCO Title: Fusion Transcript Detection - ChimeraScan

BCO Generation Date: February 16, 2023

BCO Specification Version: v1.3.0

BCO Generator: Seven Bridges

Contents

1 BioCompute Object Domain Entries

1.1 Top Level Fields


```
["https://w3id.org/biocompute/1.4.2/",
"https://biocompute.sbgenomics.com/bco/307faa23-901f-4b2e-acaa-2774214602c7",
"cb07b721df91fbcf9fd84dc750bc821a3185a5eb18c1880c60fffff423e744fa"]
```


1.2 Provenance Domain


```
{
"name": "Fusion Transcript Detection - ChimeraScan",
"version": "1.0.0",
"review": [],
"derived_from":
"https://cgc-api.sbgenomics.com/v2/apps/phil_webster/bco-cwl-examples/fusion-transcript-detecti
"obsolete_after": "2023-02-16T00:00:00+0000",
"embargo": ["2023-02-16T00:00:00+0000",
"2023-02-16T00:00:00+0000"],
"created": "2023-02-16T00:00:00+0000",
"modified": "2023-02-16T00:00:00+0000",
"contributors": [],
"license": "https://spdx.org/licenses/CC-BY-4.0.html"
}
```


1.3 Usability Domain

```
"Fusion Transcript Detection - ChimeraScan detects and
identifies fusion transcripts from paired-end RNA-Seq data
using ChimeraScan.\n\nFusion genes or chimeras are gene
alterations resulting from chromosomal rearrangements
combining exons from genes located on the same or different
chromosomes. Fusion gene products may have a new or
different function than the two fusion partners. Fusion
transcripts are frequently found in diverse types of
```

carcinomas including breast, lung, and prostate cancers, as well as melanomas and lymphomas. Detection of known (and identification of novel) gene fusions can lead to a better understanding of the triggering mechanism and progression of carcinogenesis. Moreover, recent studies suggest that fusion gene products may represent a novel therapeutic target for the treatment of human cancers.\n\nMethod:\nThis pipeline uses ChimeraScan software package that to detect fusion genes (1). Besides the primary program that detects fusion genes, the toolkit consists of an accessory tool that prepares references for proper indexing in upstream analysis and a tool for preparing output files in HTML table format. ChimeraScan is found to exhibit strong performance with low rate of false positive fusion detections, but only accepts paired-end RNA-Seq data. Addition of Chimera (2) and Oncofuse (3) to this pipeline serves to provide additional control of detected fusion genes. Finally, graphical representation of identified fusion genes is provided by the genomic coordinate visualization tool Circos (4).\n\nInputs:\nReads (paired-end): This pipeline accepts one pair of paired-end RNA-Seq data in FASTQ format (plain text or compressed files). If reads for samples are present in multiple files, the Merge FASTQ Files Public Pipeline can be used to consolidate them before alignment. If both reads are given as a gzipped archive, one can use SBG Unpack FASTQs to unpack and set the paired-end metadata fields to 1 and 2 respectively, which is obligatory.\n\nGenome reference: FASTA file containing reference genome. For human samples we recommend ucsc.hg19.fasta.\nTranscriptome reference: Transcriptome reference file containing all known transcripts in GTF format. For human samples we recommend

human_hg19_genes_2015.gtf. NOTE: ChimeraScan Index does not handle well the GENCODE GTF releases.\n\nFalse Positive Chimeras (optional): List of known false positive chimeras. For human samples we recommend hg19_bodymap_false_positive_chimeras.txt.\n\nTools and suggested parameter settings:\n* ChimeraScan v0.4.5 - Software package for detection of gene fusions in paired-end RNA-Seq datasets (1). ChimeraScan uses Bowtie to align paired-end reads to a combined genome/transcriptome reference, aiming to discover discordant reads, predict an optimal fusion breakpoint location, and detect chimeras. Software package includes an indexing program, ChimeraScan Index, which creates the combined index from genomic reference sequences (FASTA format) and custom transcriptome reference format (UCSC GenePred format). ChimeraScan GTF to genePred has been added to this pipeline to convert GTF file in a format acceptable by ChimeraScan Index. ChimeraScan HTML Table creates an HTML page with links to detailed descriptions of the chimeric genes. Additionally, SBG ChimeraScan4Circos generates output files needed visualization tool Circos. \nParameters are set to default values based on the best practice suggested by the ChimeraScan authors.\n\nTips for reducing false positives: \nOnly consider annotated genes (unless, of course, you are looking for fusion between unannotated transcripts).\nProvide a file of known likely false positives during task execution. We recommend, for human samples, hg19_bodymap_false_positive_chimeras.txt.\n* Chimera v1.12.0 - Software package for downstream processing that accepts ChimeraScan BEDPE output and enables further filtering of detected fusion transcripts (2). This tool generates a detected and a filtered fusion

genes file, and produces the R workspace file obtained from pipeline execution. \n\nParameters that can be adjusted while filtering fusion candidates: \nmin.support (default: 10), minimal number of reads spanning a specific fusion.\nfilterList type: (default: annotated genes) Additional filtering allows predicted chimeras to be discarded based on the following criteria: \nspanning reads - if it has less spanning reads than a set value\nfusion names - if particular fusion (or gene) name is in the given list (e.g. ABL1:BCR)\nintronic - if the intronic regions are included in the fusion\nannotated genes - if the partner genes are not annotated (currently set as default)\nread through - if gene partners are the same\n* Oncofuse v1.1.0 serves for prioritization of fusions based on their oncogenic potential, i.e. the probability of being 'driver' event (3). Further, a node that converts output from Oncofuse into an html table sorted by driver probability is added.\n* Circos v0.68 - Tool for visual representation of identified fusion genes (4). Graphical visualization relies on circular layout of the genome using fixed configuration file. \n\nOutputs:\nThe chimeras.bedpe file contains information about the chromosomal regions, transcript IDs, genes, and statistics for each identified chimera. \nSortable table of detected chimeras in a user-friendly HTML page, for web browser viewing, with links to detailed descriptions of the chimeric genes. \nIndex file as a gzipped archive. NOTE: In case of subsequent runs of the pipeline with the same genome/transcriptome reference, one should reuse this file (in order to save approximately 4h needed for creating an index) and modify the pipeline accordingly (to start with Chimerascan Run). \nDetected and filtered fusion genes

files generated by Chimera tool, provided as an additional control for true fusion detection; R workspace contains all saved R objects during execution of Chimera - it is useful for further analysis and for additional details on detected fusions.\nOncofuse output with functional prediction scores (oncogenic potential) of detected fusions. \nCircos plots for visual representation of fusion genes.\n\nAdditional suggestions:\nThe execution time of this pipeline can be measured in hours, even dozens of hours for big FASTQs (say, bigger than 10 GB). Remember that only creating an index takes around 4 hours. \nFeel free to customize this pipeline by removing or adding new nodes based on goal of your study.\n\n\nReferences:\n(1) Iyer, M. K. et al. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics 27, 2903-2904 (2011)\n(2) Beccuti M, Carrara M, Cordero F, Lazzarato F, Donatelli S, Nadalin F, Policriti A and Calogero RA \"Chimera: a Bioconductor package for secondary analysis of fusion products.\" Bioinformatics, 0, pp. 3 (2014)\n(3) Mikhail Shugay, Inigo Ortiz de Mendíbil, Jose L. Vizmanos and Francisco J. Novo. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics, 29 (20): 2539-2546 (2013)\n(4) Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639-1645 (2009)"

1.4 Extension Domain


```
{
  "fhir_extension": {
    "fhir_endpoint": "",
    "fhir_version": "",
    "fhir_resources": {}
```

```
  },
  "scm_extension": {
    "scm_repository": "",
    "scm_type": "git",
    "scm_commit": "",
    "scm_path": "",
    "scm_preview": ""
  }
}
```


1.5 Description Domain


```
{
"keywords": [],
"xref": [],
"platform": [
"Seven Bridges Platform"
],
"pipeline_steps": [
{
"step_number": "1",
"name": "#ChimeraScan_Index",
"description": "The ChimeraScan Index builds a combined
index using Bowtie-1.1.2 from genomic sequence (FASTA) and
transcriptome references (UCSC GenePred format). The
required format of transcriptome reference can be made from
the GTF transcriptome reference file using the ChimeraScan
Gtf2genepred tool. Output of ChimeraScan Index is used by
the fusion finder ChimeraScan Run. \n\n\n## Inputs
###\n\n**reference** - FASTA or corresponding TAR file. If
FASTA file is used then combined index (TAR file) is formed
and execution lasts approximately one hour. Resulting TAR
file can be used in any future execution as **reference**
```

```
file (if the same **reference** FASTA file and **genes**
GTF file are to be used ) as it is already appropriately
indexed. Usage of TAR file shortens execution to couple of
minutes. \n\n**genes** - GENPRED file. It is a reference
transcriptome file and it has to be compatible with the
**reference**. If GenePred file is obtained from GTF file
(using ChimeraScan Gtf2genepred tool) then GTF has to
correspond to the provided **reference**.\nExample:
human\\_hg19\\_genes\\_2015.gtf is compatible with
ucsc.hg19.fasta.\n\n## Output ###\n\n**index** - TAR file,
used further in ChimeraScan Run tool.\n\n## Common Issues
###\nEven if **genes** file does not correspond to the used
genome build (**reference** file) tool will not necessarily
break.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "2",
"name": "#SBG_Text2Html",
"description": "This is a simple R script that takes a TEXT
file and converts it into an HTML file. It uses just one
function from 'sjPlot' library; it allows sorting of rows
based on a chosen column given its name (header) or index
number.\n\n### Inputs\n**text_file** - file to be converted
to HTML\n\n**sortcolumn** - chose column for sorting
\n\n### Common issues\nIf **sortcolumn** is misspelled task
will faill. Configuration input **sortcolumn** is case
sensitive, provided string has to mach column one chooses,
other way task will fail.",
```

```
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "3",
"name": "#Oncofuse",
"description": "This tool predicts the oncogenic potential
of fusion genes found by NGS in cancer cells. It is a
post-processing step that tries to validate in-silico the
predictions made by fusion detection software. Oncofuse is
NOT a fusion detection software: its goal is NOT to
identify fusion sequences but to assign a functional
prediction score (oncogenic potential, for instance, the
probability of being 'driver' event) to fusion sequences
identified by certain fusion finder. Oncofuse is a naive
bayesian classifier built using information from Shugay et
al. 2012 and is described in Shugay et al.
2013.\n\nOncofuse can directly validate fusions obtained by
following fusion finders (input\\_type): Tophat-fusion
(tophat), FusionCatcher software (fcathcer), RNASTAR
(rnastar), STAR-Fusion (starfusion). Beside a .txt file
that contains fusions (but does not contain tissue of
origin), input_type is required input as well as
tissue\\_type. There are four pre-built libraries,
corresponding to the four supported tissue types: EPI
(epithelial origin), HEM (hematological origin), MES
(mesenchymal origin) and AVG (average expression, if tissue
source is unknown). \n\nOncofuse can be also used for
processing outputs of fusion finders not listed above. For
this purpose provided input file has to be tab-delimited
```

file with lines containing 5' and 3' breakpoint positions (first nucleotide lost upon fusion) and tissue of origin. You can find more about appropriate file format at http://www.unav.es/genetica/oncofuse.html. It is necessary to set \"coord\" as input_type format and to set \"-\" as tissue type (as tissue of origin is already included in a file). File parser (SBG Bedpe4Oncofuse) for ChimneraScan tool fusion finder can be found on our Platform and used prior to Oncofuse.\n\n###Common Issues###\nIf one of the listed fusion finders is set as input_type, origin tissue type (tissue_type) has to be set to EPI, HEM, MES or AVG, otherwise task will fail. \nIf input_type is set to \"coord\", but tissue_type is set to EPI, HEM, MES or AVG instead of \"-\" task will fail.\n\nPaper:\nMikhail Shugay, Inigo Ortiz de Mendíbil, Jose L. Vizmanos and Francisco J. Novo. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics. 16 Aug 2013. doi:10.1093/bioinformatics/btt445.",
"version": "1.1.1",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "4",
"name": "#ChimeraScan_Gtf2Genepred",
"description": "ChimeraScan GTF2GenePred is a tool that parses standard gene annotation file format (GTF) to genePred text format acceptable by ChimeraScan Index tool. It is a part of ChimeraScan package that besides ChimeraScan GTF2Genepred contains ChimeraScan Index and

ChimeraScan Run tool.\n\n#### Inputs\n**genes** - Gene feature file (GTF format) to be converted to UCSC genePred text format.\n\n### Common issues \n For this tool to work properly attribute field _gene_name_ has to be present in the GTF file.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "5",
"name": "#ChimeraScan_Make_Html",
"description": "The ChimeraScan Make HTML creates a table in the user-friendly HTML format for web browser viewing. It accepts a tab-delimited text file containing detected chimera information.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "6",
"name": "#Chimera",
"description": "Chimera is a software package for the secondary analysis of fusion products. This package facilitates the characterization of fusion products events. It allows fusion data results to be imported from the following fusion finders: ChimeraScan, bellerophontes, deFuse, FusionFinder, FusionHunter, mapSplice, tophat-fusion, FusionMap, STAR, Rsubread, and

```
fusionCatcher.\n\nChimera generates a list of detected and
filtered fusion gene files. Additionally, it can generate
files required for graphical representation of fusions with
Circos.\n\nRequired inputs are:\n\n**fusion_data** (Fusion
Data) - generated as the output of fusion finders ( bedpe,
junction, tsv...)\n\n**fusionfinder** (The Fusion Finder
Tool) - Here one has to specify tool that is used for
fusion detection, the one that generated
fusion_file.\n\n**organism** (The organism to be used for
annotation) - this is a version of a reference genome used
by Chimera tool for annotation. One can chose between hg19
and hg38. It is important that the genome reference version
used for the alignment in a fusion finder is the same of
the one used by Chimera for annotation because between hg38
and hg19 there are shifts in gene
location.\n\n**filterlist** (FilterList type) - A function
that filters out the fusion list. A fusion is discarded:
\n(i) if it has less spanning reads than a set value,
\n(ii) if its name is not in the given list, \n(iii) if the
intronic regions are included in the fusion, \n(iv) if the
partner genes are not annotated or \n(v) if gene partners
are the same, respectively.\n\n**minsupport** (Define
detected fusions by minimum supporting reads) - Parameter
\"min.support\" allows to retrieve only a subset of fusions
supported by a user defined minimal number of junction
spanning reads. If one defines a less stringent number of
supports, e.g. 2-3, more fusions supported by defined
spanning reads will be detected, normally those with low
overall quality. \n\n**filterfusionnames** (Filter detected
fusions: by fusion partner) - Search detected fusions when
fusion.names is selected in \"Filterlist type\" by
gene/fusion name or its part\n\n**filterminsupport**
```

(Filter detected fusions: by minimum supporting reads) -
Minimum number of supporting reads for the fusion not the
be filtered out applied when spanning.reads is selected as
\"filterlist\" type.\n\nPaper:\nBeccuti M, Carrara M,
Cordero F, Lazzarato F, Donatelli S, Nadalin F, Policriti A
and Calogero RA (2014). "Chimera: a Bioconductor package
for secondary analysis of fusion products." Bioinformatics,
0, pp. 3. http://doi.org/10.1093/bioinformatics/btu662.",
"version": "1.12.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "7",
"name": "#Circos",
"description": "Circos is a software package for
visualizing data and information. It applies the circular
ideogram layout to display of relationships between genomic
intervals. One timely application of this approach is
creating effective figures showing how cancer genomes
differ from healthy ones (e.g.
http://cancer.sanger.ac.uk/cosmic).\n\nIn general Circos is
ideal for exploring relationships between objects or
positions, but version hosted here is adapted only for
plotting **human fusion genes**. \n\nRequired inputs are
GFF-style data files and Apache-like configuration
files.\n\nOutput images are given in PDF.\n\nIt is used in
Fusion Transcript Detection - ChimeraScan workflow as a
tool for visualization of results obtained with Chimera
tool and with Oncofuse. Please note that SBG
Oncofuse4Circos is used for parsing Oncofuse output file

for Circos.\n\n### Inputs ###\n\n**circos_links** - input file with listed chromosomes, start and end position of each gene partner as well as its gene's name. \n\n**circos_names** - input file with both fusion partners listed with origin information (chromosome, region on a given chromosome) and gene's name of each fusion partner.\n\n### Common Issues ###\n\nTo form a proper display, Circos tool requires both files with gene names and fusion links as well as setting karyotype that is used for creation of input files.",
"version": "0.69-4",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "8",
"name": "#SBG_Bedpe4Oncofuse",
"description": "SBG Bedpe4Oncofuse is a simple one-liner that prepares ChimeraScan BEDPE output for Oncofuse analysis. There are four pre-built libraries, corresponding to the tissue types that are supported by Oncofuse : EPI (epithelial origin), HEM (hematological origin), MES (mesenchymal origin) and AVG (average expression, if tissue source is unknown) and this parameter has to be set for Oncofuse to work properly with ChimeraScan resluts.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{

```
"step_number": "9",
"name": "#Circos_1",
"description": "Circos is a software package for
visualizing data and information. It applies the circular
ideogram layout to display of relationships between genomic
intervals. One timely application of this approach is
creating effective figures showing how cancer genomes
differ from healthy ones (e.g.
http://cancer.sanger.ac.uk/cosmic).\n\nIn general Circos is
ideal for exploring relationships between objects or
positions, but version hosted here is adapted only for
plotting **human fusion genes**. \n\nRequired inputs are
GFF-style data files and Apache-like configuration
files.\n\nOutput images are given in PDF.\n\nIt is used in
Fusion Transcript Detection - ChimeraScan workflow as a
tool for visualization of results obtained with Chimera
tool and with Oncofuse. Please note that SBG
Oncofuse4Circos is used for parsing Oncofuse output file
for Circos.\n\n### Inputs ###\n\n**circos_links** - input
file with listed chromosomes, start and end position of
each gene partner as well as its gene's name.
\n\n**circos_names** - input file with both fusion partners
listed with origin information (chromosome, region on a
given chromosome) and gene's name of each fusion
partner.\n\n### Common Issues ###\n\nTo form a proper
display, Circos tool requires both files with gene names
and fusion links as well as setting karyotype that is used
for creation of input files.",
"version": "0.69-4",
"prerequisite": [],
"input_list": [],
"output_list": []
```

```
},
{
"step_number": "10",
"name": "#SBG_Html2b64",
"description": "Tool for converting HTML reports of FastQC,
SnpEff, MultiQC (simple report only) and ChimeraScan to
b64html so it can easily be displayed on SBG platform.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "11",
"name": "#SBG_Compressor",
"description": "SBG Compressor performs the
archiving(and/or compression) of the files provided on the
input. The format of the output can be selected.
\n\tSupported formats are:\n\t\t1. TAR\n\t\t2. TAR.GZ
\n\t\t3. TAR.BZ2\n\t\t4. GZ\n\t\t5. BZ2\n\t\t6. ZIP\nFor
formats TAR, TAR.GZ, TAR.BAZ2 and ZIP, a single archive
will be created on the output. For formats GZ and BZ2, one
archive per file will be created.",
"version": "v1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "12",
"name": "#SBG_Oncofuse4Circos",
"description": "An R script that extracts fusion links and
```

```
names needed from the Oncofuse output and prepares it for
Circos tool.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "13",
"name": "#ChimeraScan_Run",
"description": "ChimeraScan detects fusion genes (chimeras)
in paired-end RNA-seq datasets. This tool uses the Bowtie
aligner to align paired-end reads to a combined
genome-transcriptome reference. It aims to discover
discordant reads, predict an optimal fusion breakpoint
location, and detect chimeras.This application outputs a
tabular file (*.chimeras.bedpe) that contains information
about the chromosomal regions, transcript IDs, genes, and
statistics for each chimera.\n\n### Inputs\n\n**reads** -
RNA-Seq FASTQ paired-end files.\n\n**index** - TAR file
created by ChimeraScan Index tool.  \n\n**false_positives**
- TXT supporting file containing list of likely false
positives (hg19 Homo Sapiens),
https://code.google.com/archive/p/chimerascan/downloads\n\n###
Common issues:\n\nThe paired-end reads must be of the same
length.\n\nFASTQ.GZ files provided as **reads** instead of
FASTQ.\n \n\nReferences:\n\n1. Maher, C.A., et al.
Transcriptome sequencing to detect gene fusions in cancer.
Nature 458, 97-101 (2009).\n\n2. Maher, C.A., et al.
Chimeric transcript discovery by paired-end transcriptome
sequencing. Proceedings of the National Academy of Sciences
of the United States of America 106, 12353-12358 (2009).",
```

```
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
}
]
}
```


1.6 Execution Domain


```
{
"keywords": [],
"xref": [],
"platform": [
"Seven Bridges Platform"
],
"pipeline_steps": [
{
"step_number": "1",
"name": "#ChimeraScan_Index",
"description": "The ChimeraScan Index builds a combined
index using Bowtie-1.1.2 from genomic sequence (FASTA) and
transcriptome references (UCSC GenePred format). The
required format of transcriptome reference can be made from
the GTF transcriptome reference file using the ChimeraScan
Gtf2genepred tool. Output of ChimeraScan Index is used by
the fusion finder ChimeraScan Run. \n\n\n## Inputs
###\n\n**reference** - FASTA or corresponding TAR file. If
FASTA file is used then combined index (TAR file) is formed
and execution lasts approximately one hour. Resulting TAR
file can be used in any future execution as **reference**
file (if the same **reference** FASTA file and **genes**
GTF file are to be used ) as it is already appropriately
```

indexed. Usage of TAR file shortens execution to couple of minutes. \n\n**genes** - GENPRED file. It is a reference transcriptome file and it has to be compatible with the **reference**. If GenePred file is obtained from GTF file (using ChimeraScan Gtf2genepred tool) then GTF has to correspond to the provided **reference**.\nExample: human_hg19_genes_2015.gtf is compatible with ucsc.hg19.fasta.\n\n## Output ###\n\n**index** - TAR file, used further in ChimeraScan Run tool.\n\n## Common Issues ###\nEven if **genes** file does not correspond to the used genome build (**reference** file) tool will not necessarily break.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "2",
"name": "#SBG_Text2Html",
"description": "This is a simple R script that takes a TEXT file and converts it into an HTML file. It uses just one function from 'sjPlot' library; it allows sorting of rows based on a chosen column given its name (header) or index number.\n\n### Inputs\n**text_file** - file to be converted to HTML\n\n**sortcolumn** - chose column for sorting \n\n### Common issues\nIf **sortcolumn** is misspelled task will faill. Configuration input **sortcolumn** is case sensitive, provided string has to mach column one chooses, other way task will fail.",
"version": "1.0",
"prerequisite": [],

```
"input_list": [],
"output_list": []
},
{
"step_number": "3",
"name": "#Oncofuse",
"description": "This tool predicts the oncogenic potential
of fusion genes found by NGS in cancer cells. It is a
post-processing step that tries to validate in-silico the
predictions made by fusion detection software. Oncofuse is
NOT a fusion detection software: its goal is NOT to
identify fusion sequences but to assign a functional
prediction score (oncogenic potential, for instance, the
probability of being 'driver' event) to fusion sequences
identified by certain fusion finder. Oncofuse is a naive
bayesian classifier built using information from Shugay et
al. 2012 and is described in Shugay et al.
2013.\n\nOncofuse can directly validate fusions obtained by
following fusion finders (input\\_type): Tophat-fusion
(tophat), FusionCatcher software (fcathcer), RNASTAR
(rnastar), STAR-Fusion (starfusion). Beside a .txt file
that contains fusions (but does not contain tissue of
origin), input_type is required input as well as
tissue\\_type. There are four pre-built libraries,
corresponding to the four supported tissue types: EPI
(epithelial origin), HEM (hematological origin), MES
(mesenchymal origin) and AVG (average expression, if tissue
source is unknown). \n\nOncofuse can be also used for
processing outputs of fusion finders not listed above. For
this purpose provided input file has to be tab-delimited
file with lines containing 5' and 3' breakpoint positions
(first nucleotide lost upon fusion) and tissue of origin.
```

You can find more about appropriate file format at http://www.unav.es/genetica/oncofuse.html. It is necessary to set \"coord\" as input_type format and to set \"-\" as tissue type (as tissue of origin is already included in a file). File parser (SBG Bedpe4Oncofuse) for ChimneraScan tool fusion finder can be found on our Platform and used prior to Oncofuse.\n\n###Common Issues###\nIf one of the listed fusion finders is set as input_type, origin tissue type (tissue_type) has to be set to EPI, HEM, MES or AVG, otherwise task will fail. \nIf input_type is set to \"coord\", but tissue_type is set to EPI, HEM, MES or AVG instead of \"-\" task will fail.\n\nPaper:\nMikhail Shugay, Inigo Ortiz de Mendíbil, Jose L. Vizmanos and Francisco J. Novo. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics. 16 Aug 2013. doi:10.1093/bioinformatics/btt445.",
"version": "1.1.1",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "4",
"name": "#ChimeraScan_Gtf2Genepred",
"description": "ChimeraScan GTF2GenePred is a tool that parses standard gene annotation file format (GTF) to genePred text format acceptable by ChimeraScan Index tool. It is a part of ChimeraScan package that besides ChimeraScan GTF2Genepred contains ChimeraScan Index and ChimeraScan Run tool.\n\n#### Inputs\n**genes** - Gene feature file (GTF format) to be converted to UCSC genePred

text format.\n\n### Common issues \n For this tool to work properly attribute field _gene_name_ has to be present in the GTF file.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "5",
"name": "#ChimeraScan_Make_Html",
"description": "The ChimeraScan Make HTML creates a table in the user-friendly HTML format for web browser viewing. It accepts a tab-delimited text file containing detected chimera information.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "6",
"name": "#Chimera",
"description": "Chimera is a software package for the secondary analysis of fusion products. This package facilitates the characterization of fusion products events. It allows fusion data results to be imported from the following fusion finders: ChimeraScan, bellerophontes, deFuse, FusionFinder, FusionHunter, mapSplice, tophat-fusion, FusionMap, STAR, Rsubread, and fusionCatcher.\n\nChimera generates a list of detected and filtered fusion gene files. Additionally, it can generate

```
files required for graphical representation of fusions with
Circos.\n\nRequired inputs are:\n\n**fusion_data** (Fusion
Data) - generated as the output of fusion finders ( bedpe,
junction, tsv...)\n\n**fusionfinder** (The Fusion Finder
Tool) - Here one has to specify tool that is used for
fusion detection, the one that generated
fusion_file.\n\n**organism** (The organism to be used for
annotation) - this is a version of a reference genome used
by Chimera tool for annotation. One can chose between hg19
and hg38. It is important that the genome reference version
used for the alignment in a fusion finder is the same of
the one used by Chimera for annotation because between hg38
and hg19 there are shifts in gene
location.\n\n**filterlist** (FilterList type) - A function
that filters out the fusion list. A fusion is discarded:
\n(i) if it has less spanning reads than a set value,
\n(ii) if its name is not in the given list, \n(iii) if the
intronic regions are included in the fusion, \n(iv) if the
partner genes are not annotated or \n(v) if gene partners
are the same, respectively.\n\n**minsupport** (Define
detected fusions by minimum supporting reads) - Parameter
\"min.support\" allows to retrieve only a subset of fusions
supported by a user defined minimal number of junction
spanning reads. If one defines a less stringent number of
supports, e.g. 2-3, more fusions supported by defined
spanning reads will be detected, normally those with low
overall quality. \n\n**filterfusionnames** (Filter detected
fusions: by fusion partner) - Search detected fusions when
fusion.names is selected in \"Filterlist type\" by
gene/fusion name or its part\n\n**filterminsupport**
(Filter detected fusions: by minimum supporting reads) -
Minimum number of supporting reads for the fusion not the
```

be filtered out applied when spanning.reads is selected as \"filterlist\" type.\n\nPaper:\nBeccuti M, Carrara M, Cordero F, Lazzarato F, Donatelli S, Nadalin F, Policriti A and Calogero RA (2014). "Chimera: a Bioconductor package for secondary analysis of fusion products." Bioinformatics, 0, pp. 3. http://doi.org/10.1093/bioinformatics/btu662.",
"version": "1.12.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "7",
"name": "#Circos",
"description": "Circos is a software package for visualizing data and information. It applies the circular ideogram layout to display of relationships between genomic intervals. One timely application of this approach is creating effective figures showing how cancer genomes differ from healthy ones (e.g. http://cancer.sanger.ac.uk/cosmic).\n\nIn general Circos is ideal for exploring relationships between objects or positions, but version hosted here is adapted only for plotting **human fusion genes**. \n\nRequired inputs are GFF-style data files and Apache-like configuration files.\n\nOutput images are given in PDF.\n\nIt is used in Fusion Transcript Detection - ChimeraScan workflow as a tool for visualization of results obtained with Chimera tool and with Oncofuse. Please note that SBG Oncofuse4Circos is used for parsing Oncofuse output file for Circos.\n\n### Inputs ###\n\n**circos_links** - input file with listed chromosomes, start and end position of

each gene partner as well as its gene's name.
\n\n**circos_names** - input file with both fusion partners
listed with origin information (chromosome, region on a
given chromosome) and gene's name of each fusion
partner.\n\n### Common Issues ###\n\nTo form a proper
display, Circos tool requires both files with gene names
and fusion links as well as setting karyotype that is used
for creation of input files.",
"version": "0.69-4",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "8",
"name": "#SBG_Bedpe4Oncofuse",
"description": "SBG Bedpe4Oncofuse is a simple one-liner
that prepares ChimeraScan BEDPE output for Oncofuse
analysis. There are four pre-built libraries, corresponding
to the tissue types that are supported by Oncofuse : EPI
(epithelial origin), HEM (hematological origin), MES
(mesenchymal origin) and AVG (average expression, if tissue
source is unknown) and this parameter has to be set for
Oncofuse to work properly with ChimeraScan resluts.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "9",
"name": "#Circos_1",

"description": "Circos is a software package for visualizing data and information. It applies the circular ideogram layout to display of relationships between genomic intervals. One timely application of this approach is creating effective figures showing how cancer genomes differ from healthy ones (e.g. http://cancer.sanger.ac.uk/cosmic).\n\nIn general Circos is ideal for exploring relationships between objects or positions, but version hosted here is adapted only for plotting **human fusion genes**. \n\nRequired inputs are GFF-style data files and Apache-like configuration files.\n\nOutput images are given in PDF.\n\nIt is used in Fusion Transcript Detection - ChimeraScan workflow as a tool for visualization of results obtained with Chimera tool and with Oncofuse. Please note that SBG Oncofuse4Circos is used for parsing Oncofuse output file for Circos.\n\n### Inputs ###\n\n**circos_links** - input file with listed chromosomes, start and end position of each gene partner as well as its gene's name. \n\n**circos_names** - input file with both fusion partners listed with origin information (chromosome, region on a given chromosome) and gene's name of each fusion partner.\n\n### Common Issues ###\n\nTo form a proper display, Circos tool requires both files with gene names and fusion links as well as setting karyotype that is used for creation of input files.",
"version": "0.69-4",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{

```
"step_number": "10",
"name": "#SBG_Html2b64",
"description": "Tool for converting HTML reports of FastQC,
SnpEff, MultiQC (simple report only) and ChimeraScan to
b64html so it can easily be displayed on SBG platform.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "11",
"name": "#SBG_Compressor",
"description": "SBG Compressor performs the
archiving(and/or compression) of the files provided on the
input. The format of the output can be selected.
\n\tSupported formats are:\n\t\t1. TAR\n\t\t2. TAR.GZ
\n\t\t3. TAR.BZ2\n\t\t4. GZ\n\t\t5. BZ2\n\t\t6. ZIP\nFor
formats TAR, TAR.GZ, TAR.BAZ2 and ZIP, a single archive
will be created on the output. For formats GZ and BZ2, one
archive per file will be created.",
"version": "v1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "12",
"name": "#SBG_Oncofuse4Circos",
"description": "An R script that extracts fusion links and
names needed from the Oncofuse output and prepares it for
Circos tool.",
```

```
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "13",
"name": "#ChimeraScan_Run",
"description": "ChimeraScan detects fusion genes (chimeras)
in paired-end RNA-seq datasets. This tool uses the Bowtie
aligner to align paired-end reads to a combined
genome-transcriptome reference. It aims to discover
discordant reads, predict an optimal fusion breakpoint
location, and detect chimeras.This application outputs a
tabular file (*.chimeras.bedpe) that contains information
about the chromosomal regions, transcript IDs, genes, and
statistics for each chimera.\n\n### Inputs\n\n**reads** -
RNA-Seq FASTQ paired-end files.\n\n**index** - TAR file
created by ChimeraScan Index tool.  \n\n**false_positives**
- TXT supporting file containing list of likely false
positives (hg19 Homo Sapiens),
https://code.google.com/archive/p/chimerascan/downloads\n\n###
Common issues:\n\nThe paired-end reads must be of the same
length.\n\nFASTQ.GZ files provided as **reads** instead of
FASTQ.\n \n\nReferences:\n\n1. Maher, C.A., et al.
Transcriptome sequencing to detect gene fusions in cancer.
Nature 458, 97-101 (2009).\n\n2. Maher, C.A., et al.
Chimeric transcript discovery by paired-end transcriptome
sequencing. Proceedings of the National Academy of Sciences
of the United States of America 106, 12353-12358 (2009).",
"version": "0.4.5",
"prerequisite": [],
```

```
"input_list": [],
"output_list": []
}
]
}
```


1.7 Parametric Domain


```
{
"keywords": [],
"xref": [],
"platform": [
"Seven Bridges Platform"
],
"pipeline_steps": [
{
"step_number": "1",
"name": "#ChimeraScan_Index",
"description": "The ChimeraScan Index builds a combined
index using Bowtie-1.1.2 from genomic sequence (FASTA) and
transcriptome references (UCSC GenePred format). The
required format of transcriptome reference can be made from
the GTF transcriptome reference file using the ChimeraScan
Gtf2genepred tool. Output of ChimeraScan Index is used by
the fusion finder ChimeraScan Run. \n\n\n## Inputs
###\n\n**reference** - FASTA or corresponding TAR file. If
FASTA file is used then combined index (TAR file) is formed
and execution lasts approximately one hour. Resulting TAR
file can be used in any future execution as **reference**
file (if the same **reference** FASTA file and **genes**
GTF file are to be used ) as it is already appropriately
indexed. Usage of TAR file shortens execution to couple of
minutes. \n\n**genes** - GENPRED file. It is a reference
```

transcriptome file and it has to be compatible with the **reference**. If GenePred file is obtained from GTF file (using ChimeraScan Gtf2genepred tool) then GTF has to correspond to the provided **reference**.\nExample: human_hg19_genes_2015.gtf is compatible with ucsc.hg19.fasta.\n\n## Output ###\n\n**index** - TAR file, used further in ChimeraScan Run tool.\n\n## Common Issues ###\nEven if **genes** file does not correspond to the used genome build (**reference** file) tool will not necessarily break.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "2",
"name": "#SBG_Text2Html",
"description": "This is a simple R script that takes a TEXT file and converts it into an HTML file. It uses just one function from 'sjPlot' library; it allows sorting of rows based on a chosen column given its name (header) or index number.\n\n### Inputs\n**text_file** - file to be converted to HTML\n\n**sortcolumn** - chose column for sorting \n\n### Common issues\nIf **sortcolumn** is misspelled task will faill. Configuration input **sortcolumn** is case sensitive, provided string has to mach column one chooses, other way task will fail.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []

},
{
"step_number": "3",
"name": "#Oncofuse",
"description": "This tool predicts the oncogenic potential
of fusion genes found by NGS in cancer cells. It is a
post-processing step that tries to validate in-silico the
predictions made by fusion detection software. Oncofuse is
NOT a fusion detection software: its goal is NOT to
identify fusion sequences but to assign a functional
prediction score (oncogenic potential, for instance, the
probability of being 'driver' event) to fusion sequences
identified by certain fusion finder. Oncofuse is a naive
bayesian classifier built using information from Shugay et
al. 2012 and is described in Shugay et al.
2013.\n\nOncofuse can directly validate fusions obtained by
following fusion finders (input_type): Tophat-fusion
(tophat), FusionCatcher software (fcathcer), RNASTAR
(rnastar), STAR-Fusion (starfusion). Beside a .txt file
that contains fusions (but does not contain tissue of
origin), input_type is required input as well as
tissue_type. There are four pre-built libraries,
corresponding to the four supported tissue types: EPI
(epithelial origin), HEM (hematological origin), MES
(mesenchymal origin) and AVG (average expression, if tissue
source is unknown). \n\nOncofuse can be also used for
processing outputs of fusion finders not listed above. For
this purpose provided input file has to be tab-delimited
file with lines containing 5' and 3' breakpoint positions
(first nucleotide lost upon fusion) and tissue of origin.
You can find more about appropriate file format at
http://www.unav.es/genetica/oncofuse.html. It is necessary

to set \"coord\" as input_type format and to set \"-\" as tissue type (as tissue of origin is already included in a file). File parser (SBG Bedpe4Oncofuse) for ChimneraScan tool fusion finder can be found on our Platform and used prior to Oncofuse.\n\n###Common Issues###\nIf one of the listed fusion finders is set as input_type, origin tissue type (tissue_type) has to be set to EPI, HEM, MES or AVG, otherwise task will fail. \nIf input_type is set to \"coord\", but tissue_type is set to EPI, HEM, MES or AVG instead of \"-\" task will fail.\n\nPaper:\nMikhail Shugay, Inigo Ortiz de Mendíbil, Jose L. Vizmanos and Francisco J. Novo. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics. 16 Aug 2013. doi:10.1093/bioinformatics/btt445.",
"version": "1.1.1",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "4",
"name": "#ChimeraScan_Gtf2Genepred",
"description": "ChimeraScan GTF2GenePred is a tool that parses standard gene annotation file format (GTF) to genePred text format acceptable by ChimeraScan Index tool. It is a part of ChimeraScan package that besides ChimeraScan GTF2Genepred contains ChimeraScan Index and ChimeraScan Run tool.\n\n#### Inputs\n**genes** - Gene feature file (GTF format) to be converted to UCSC genePred text format.\n\n### Common issues \n For this tool to work properly attribute field _gene_name_ has to be present in

```
the GTF file.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "5",
"name": "#ChimeraScan_Make_Html",
"description": "The ChimeraScan Make HTML creates a table
in the user-friendly HTML format for web browser viewing.
It accepts a tab-delimited text file containing detected
chimera information.",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "6",
"name": "#Chimera",
"description": "Chimera is a software package for the
secondary analysis of fusion products. This package
facilitates the characterization of fusion products events.
It allows fusion data results to be imported from the
following fusion finders: ChimeraScan, bellerophontes,
deFuse, FusionFinder, FusionHunter, mapSplice,
tophat-fusion, FusionMap, STAR, Rsubread, and
fusionCatcher.\n\nChimera generates a list of detected and
filtered fusion gene files. Additionally, it can generate
files required for graphical representation of fusions with
Circos.\n\nRequired inputs are:\n\n**fusion_data** (Fusion
```

```
Data) - generated as the output of fusion finders ( bedpe,
junction, tsv...)\n\n**fusionfinder** (The Fusion Finder
Tool) - Here one has to specify tool that is used for
fusion detection, the one that generated
fusion_file.\n\n**organism** (The organism to be used for
annotation) - this is a version of a reference genome used
by Chimera tool for annotation. One can chose between hg19
and hg38. It is important that the genome reference version
used for the alignment in a fusion finder is the same of
the one used by Chimera for annotation because between hg38
and hg19 there are shifts in gene
location.\n\n**filterlist** (FilterList type) - A function
that filters out the fusion list. A fusion is discarded:
\n(i) if it has less spanning reads than a set value,
\n(ii) if its name is not in the given list, \n(iii) if the
intronic regions are included in the fusion, \n(iv) if the
partner genes are not annotated or \n(v) if gene partners
are the same, respectively.\n\n**minsupport** (Define
detected fusions by minimum supporting reads) - Parameter
\"min.support\" allows to retrieve only a subset of fusions
supported by a user defined minimal number of junction
spanning reads. If one defines a less stringent number of
supports, e.g. 2-3, more fusions supported by defined
spanning reads will be detected, normally those with low
overall quality. \n\n**filterfusionnames** (Filter detected
fusions: by fusion partner) - Search detected fusions when
fusion.names is selected in \"Filterlist type\" by
gene/fusion name or its part\n\n**filterminsupport**
(Filter detected fusions: by minimum supporting reads) -
Minimum number of supporting reads for the fusion not the
be filtered out applied when spanning.reads is selected as
\"filterlist\" type.\n\nPaper:\nBeccuti M, Carrara M,
```

Cordero F, Lazzarato F, Donatelli S, Nadalin F, Policriti A and Calogero RA (2014). "Chimera: a Bioconductor package for secondary analysis of fusion products." Bioinformatics, 0, pp. 3. http://doi.org/10.1093/bioinformatics/btu662.",
"version": "1.12.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "7",
"name": "#Circos",
"description": "Circos is a software package for visualizing data and information. It applies the circular ideogram layout to display of relationships between genomic intervals. One timely application of this approach is creating effective figures showing how cancer genomes differ from healthy ones (e.g. http://cancer.sanger.ac.uk/cosmic).\n\nIn general Circos is ideal for exploring relationships between objects or positions, but version hosted here is adapted only for plotting **human fusion genes**. \n\nRequired inputs are GFF-style data files and Apache-like configuration files.\n\nOutput images are given in PDF.\n\nIt is used in Fusion Transcript Detection - ChimeraScan workflow as a tool for visualization of results obtained with Chimera tool and with Oncofuse. Please note that SBG Oncofuse4Circos is used for parsing Oncofuse output file for Circos.\n\n### Inputs ###\n\n**circos_links** - input file with listed chromosomes, start and end position of each gene partner as well as its gene's name. \n\n**circos_names** - input file with both fusion partners

listed with origin information (chromosome, region on a
given chromosome) and gene's name of each fusion
partner.\n\n### Common Issues ###\n\nTo form a proper
display, Circos tool requires both files with gene names
and fusion links as well as setting karyotype that is used
for creation of input files.",
"version": "0.69-4",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "8",
"name": "#SBG_Bedpe4Oncofuse",
"description": "SBG Bedpe4Oncofuse is a simple one-liner
that prepares ChimeraScan BEDPE output for Oncofuse
analysis. There are four pre-built libraries, corresponding
to the tissue types that are supported by Oncofuse : EPI
(epithelial origin), HEM (hematological origin), MES
(mesenchymal origin) and AVG (average expression, if tissue
source is unknown) and this parameter has to be set for
Oncofuse to work properly with ChimeraScan resluts.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "9",
"name": "#Circos_1",
"description": "Circos is a software package for
visualizing data and information. It applies the circular

ideogram layout to display of relationships between genomic intervals. One timely application of this approach is creating effective figures showing how cancer genomes differ from healthy ones (e.g. http://cancer.sanger.ac.uk/cosmic).\n\nIn general Circos is ideal for exploring relationships between objects or positions, but version hosted here is adapted only for plotting **human fusion genes**. \n\nRequired inputs are GFF-style data files and Apache-like configuration files.\n\nOutput images are given in PDF.\n\nIt is used in Fusion Transcript Detection - ChimeraScan workflow as a tool for visualization of results obtained with Chimera tool and with Oncofuse. Please note that SBG Oncofuse4Circos is used for parsing Oncofuse output file for Circos.\n\n### Inputs ###\n\n**circos_links** - input file with listed chromosomes, start and end position of each gene partner as well as its gene's name. \n\n**circos_names** - input file with both fusion partners listed with origin information (chromosome, region on a given chromosome) and gene's name of each fusion partner.\n\n### Common Issues ###\n\nTo form a proper display, Circos tool requires both files with gene names and fusion links as well as setting karyotype that is used for creation of input files.",
"version": "0.69-4",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "10",
"name": "#SBG_Html2b64",

```
"description": "Tool for converting HTML reports of FastQC,
SnpEff, MultiQC (simple report only) and ChimeraScan to
b64html so it can easily be displayed on SBG platform.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "11",
"name": "#SBG_Compressor",
"description": "SBG Compressor performs the
archiving(and/or compression) of the files provided on the
input. The format of the output can be selected.
\n\tSupported formats are:\n\t\t1. TAR\n\t\t2. TAR.GZ
\n\t\t3. TAR.BZ2\n\t\t4. GZ\n\t\t5. BZ2\n\t\t6. ZIP\nFor
formats TAR, TAR.GZ, TAR.BAZ2 and ZIP, a single archive
will be created on the output. For formats GZ and BZ2, one
archive per file will be created.",
"version": "v1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "12",
"name": "#SBG_Oncofuse4Circos",
"description": "An R script that extracts fusion links and
names needed from the Oncofuse output and prepares it for
Circos tool.",
"version": "1.0",
"prerequisite": [],
```

"input_list": [],
"output_list": []
},
{
"step_number": "13",
"name": "#ChimeraScan_Run",
"description": "ChimeraScan detects fusion genes (chimeras)
in paired-end RNA-seq datasets. This tool uses the Bowtie
aligner to align paired-end reads to a combined
genome-transcriptome reference. It aims to discover
discordant reads, predict an optimal fusion breakpoint
location, and detect chimeras.This application outputs a
tabular file (*.chimeras.bedpe) that contains information
about the chromosomal regions, transcript IDs, genes, and
statistics for each chimera.\n\n### Inputs\n\n**reads** -
RNA-Seq FASTQ paired-end files.\n\n**index** - TAR file
created by ChimeraScan Index tool. \n\n**false_positives**
- TXT supporting file containing list of likely false
positives (hg19 Homo Sapiens),
https://code.google.com/archive/p/chimerascan/downloads\n\n###
Common issues:\n\nThe paired-end reads must be of the same
length.\n\nFASTQ.GZ files provided as **reads** instead of
FASTQ.\n \n\nReferences:\n\n1. Maher, C.A., et al.
Transcriptome sequencing to detect gene fusions in cancer.
Nature 458, 97-101 (2009).\n\n2. Maher, C.A., et al.
Chimeric transcript discovery by paired-end transcriptome
sequencing. Proceedings of the National Academy of Sciences
of the United States of America 106, 12353-12358 (2009).",
"version": "0.4.5",
"prerequisite": [],
"input_list": [],
"output_list": []

```
}
]
}
```

1.8 Input/Output Domain

```
{
  "input_subdomain": [
    {
      "uri": [
        {
          "filename": "",
          "uri": "",
          "access_time": ""
        }
      ]
    }
  ],
  "output_subdomain": [
    {
      "mediatype": "",
      "uri": [
        {
          "uri": "",
          "access_time": ""
        }
      ]
    }
  ]
}
```

1.9 Error Domain

```
{
```

```
  "empirical_error": [],
  "algorithmic_error": []
}
```

2 Funding

The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

3 References

Lau et al (2017) The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. Cancer Res. 77(21):e3-e6. doi: 10.1158/0008-5472.CAN-17-0387.

4 Appendix 1: BioCompute Object Specification v1.3.0

Name	ID	Description
Top Level Fields		
BioCompute Object Indetifier	BCO_id	Unique identifier that should be apllied to each BCO instance. Assigned by a BCO database engine, like URL. It never be reused.
Type	type	As any object of the type, it has its own fields.
Digital signature	digital_signature	A string-type, read-only generated and stored by a BCO database, protecting the object from internal or external alterations without proper validation. It can be used for validation, downloading, and transferring BCOs.
BCO version	bco_spec_version	The version of the BCO specification used to define this document.
Provenance Domain		
Name	name	Name of the BCO.
Structured name	structured_name	Computable text field designed to represent a BCO instance name in visible interfaces
Version	version	Records the versioning of this BCO instance object. A change in the BCO affecting the outcome of the computation should be deposited as a new BCO, not as a new version.
Review	review	Describes the status of an object in the review process. Status flags: unreviewed, in-review, approved, suspended, rejected.
Inheritance/derivation	derived_from	If the object is derived from another, this field will specify the parent object, in the form of the objectid. It is null, if inherits only from the base BioCompute Object or a type definition.
Obsolescence	obsolete	If the object has an expiration date this field will specify that using the datetime type.
Embargo	embargo	If the object has a period of time that it is not public, that range can be specified using these fields. Using the datetime type a start and end time are specified for the embargo.
Created	created	Using the datetime type the time of initial creation of the BCO is recorded.
Modification	modified	Using the datetime type the time of most recent modification of the BCO is recorded.
Contributors	contributors	List to hold contributor identifiers and a description of their type of contribution, including a field for ORCIDs to record author information, as they allow for the author to curate their information after submission.
License	license	A space for Creative commons licence or other licence information. The default or recommended licence can be Attribution 4.0 International.
Usability Domain		
Usability Domain	usability_domain	Provides a space for the author to define the usability domain of the BCO. It is an array of free text values. This field is to aid in search-ability and provide a specific description of the object. It helps determine when and how the BCO can be used.
Extension Domain		

(continued)

Name	ID	Description
Extension Domain	extension_domain	For a user to add more structured information that is defined in the type definition. This section is not evaluated by checks for BCO validity or computational correctness.
Extension to External References: SMART on FHIR Genomics	Extension to External References: SMART on FHIR Genomics	SMART on FHIR Genomics provides a framework for HER-based apps to built on FHIR that integrate clinical and genomics information.
Extension to External References: GitHub	Extension to External References: GitHub	Include an extension to GitHub repositories where HTS computational analysis pipelins, workflows, protocols, and tool or software source code can be stored, deposited, downloaded.
Description Domain		
Description Domain	description_domain	Structured field for description of external references, the pipeline steps, and the relationship of IO objects. Information in this domain is not used for computation. Capture information that is currently being provided in FDA submission in journal format.
Keywords	keywords	List of key map fields to hold a list of keywords to aid in search-ability and description of the object.
External References	xref	It contains a list of the databases and/or ontology IDs that are cross-referenced in the BCO. It provides more specificity in the information related to BCO entries.
Pipeline tools	pipeline_steps	For recording the specifics of a pipeline. Each individual tool is represented as step, at the discretion of the author. Step Number (step_number), Name (name), Tool Description (description), Tool Version (version), Tool Prerequisites (prerequisite), Input List (input_list), Output List (output_list).
Execution Domain		
Execution Domain	execution_domain	The filelds required for execution of the BCO have been encapsulated together in order to clearly separate information needed for deployment, software configuration, and running applications in a dependent enviroment.
Script Access Type	script_access_type	This field indicates whether the code of the script to execute the BioCompute Object is access as an external file via HTTP or in-line text in the script field.
Script	script	Points to an internal or external reference to a script object that was used to perform computations for this BCO instance. This may be reference to Galaxy Project or Seven Bridges Genomics pipeline, a Common Workflow Language (CWL) object in GitHub, HIVE computational service or any other type of script.
Pipeline Version	pipeline_version	This field records the version of the pipeline implementation.
Platform/Environment	platform	The multi-value reference to a particular deployment of an existing platform where this BCO can be reproduced (Galaxy or HIVE or CASAVA).
Script Driver	script_driver	The reference to an executable that can be launched in order to perform a sequence of commands described in the script. For example if the pipeline is driven by a HIVE script, the script driver is the hive execution engine. For CWL based scripts specify cwl-runner. Another very general commonly used in Linux based operating systems is shell.

(continued)

Name	ID	Description
Algorithmic tools and Software Prerequisites	software_prerequisites	Field listing the minimal necessary prerequisites, library, tool versions needed to successfully run the script to produce BCO.
Domain Prerequisites	domain_prerequisites	Listing the minimal necessary domain specific external data source access in order to successfully run the script to produce BCO.
Enviromental parameters	env_parameters	Multi-value additional key value pairs useful to configure the execution environment on the target platform, like compute cores, available memory use of the script.
Parametric Domain		
Parametric Domain	parametric_domain	List of parameters customizing the computational flow which can affect the output of the calculations. These fields are custom to each type of analysis and are tied to a particular pipeline implementation.
Input and Output Domain		
Input and output Domain	io_domain	This represents the list of global input and output files created by the computational workflow, excluding the intermediate files.
Input Subdomain	input_subdomain	This field records the references and input files for the entire pipeline. Each type of input file is listed under a key for that type.
Output Subdomain	output_subdomain	This field records the outputs for the entire pipeline .
Error Domain, acceptable range of variability		
Error Domain, acceptable range of variability	error_domain	Consists of two subdomains: empirical and algorithmic._The empirical subdomain contains the limits of_detectability_ fps, fns, statistical confidence of outcomes, etc. The algorithmic subdomain is descriptive of errors that originated by fuzziness of the algorithms, driven by stochastic processes, in dynamically parallelized multi-threaded executions, or in machine learning methodologies where the state of the machine can affect the outcomeConsists of two subdomains: empirical and algorithmic. The empirical subdomain contains the limits of detectability FPs, FNs, statistical confidence of outcomes, etc. The algorithmic subdomain is descriptive of errors that originated by fuzziness of the algorithms, driven by stochastic processes, in dynamically parallelized multi-threaded executions, or in machine learning methodologies where the state of the machine can affect the outcome.

5 Appendix 2: The Complete BioCompute Object


```
{
  "spec_version": "https://w3id.org/biocompute/1.4.2/",
  "object_id": "https://biocompute.sbgenomics.com/bco/307faa23-901f-4b2e-acaa-2774214602c7",
  "etag": "cb07b721df91fbcf9fd84dc750bc821a3185a5eb18c1880c60fffff423e744fa",
  "provenance_domain": {
    "name": "Fusion Transcript Detection - ChimeraScan",
    "version": "1.0.0",
    "review": [],
    "derived_from": "https://cgc-api.sbgenomics.com/v2/apps/phil_webster/bco-cwl-examples/fusi
    "obsolete_after": "2023-02-16T00:00:00+0000",
    "embargo": ["2023-02-16T00:00:00+0000", "2023-02-16T00:00:00+0000"],
    "created": "2023-02-16T00:00:00+0000",
    "modified": "2023-02-16T00:00:00+0000",
    "contributors": [],
    "license": "https://spdx.org/licenses/CC-BY-4.0.html"
  },
  "usability_domain": "Fusion Transcript Detection - ChimeraScan detects and identifies fusion
  "extension_domain": {
    "fhir_extension": {
      "fhir_endpoint": "",
      "fhir_version": "",
      "fhir_resources": {}
    },
    "scm_extension": {
      "scm_repository": "",
      "scm_type": "git",
      "scm_commit": "",
      "scm_path": "",
      "scm_preview": ""
    }
```

```
},
"description_domain": {
  "keywords": [],
  "xref": [],
  "platform": [
    "Seven Bridges Platform"
  ],
  "pipeline_steps": [
    {
      "step_number": "1",
      "name": "#ChimeraScan_Index",
      "description": "The ChimeraScan Index builds a combined index using Bowtie-1.1.2 from g
      "version": "0.4.5",
      "prerequisite": [],
      "input_list": [],
      "output_list": []
    },
    {
      "step_number": "2",
      "name": "#SBG_Text2Html",
      "description": "This is a simple R script that takes a TEXT file and converts it into a
      "version": "1.0",
      "prerequisite": [],
      "input_list": [],
      "output_list": []
    },
    {
      "step_number": "3",
      "name": "#Oncofuse",
      "description": "This tool predicts the oncogenic potential of fusion genes found by NGS
      "version": "1.1.1",
      "prerequisite": [],
```

```
    "input_list": [],
    "output_list": []
  },
  {
    "step_number": "4",
    "name": "#ChimeraScan_Gtf2Genepred",
    "description": "ChimeraScan GTF2GenePred is a tool that parses standard gene annotatio
    "version": "0.4.5",
    "prerequisite": [],
    "input_list": [],
    "output_list": []
  },
  {
    "step_number": "5",
    "name": "#ChimeraScan_Make_Html",
    "description": "The ChimeraScan Make HTML creates a table in the user-friendly HTML fo
    "version": "0.4.5",
    "prerequisite": [],
    "input_list": [],
    "output_list": []
  },
  {
    "step_number": "6",
    "name": "#Chimera",
    "description": "Chimera is a software package for the secondary analysis of fusion pro
    "version": "1.12.0",
    "prerequisite": [],
    "input_list": [],
    "output_list": []
  },
  {
    "step_number": "7",
```

```
    "name": "#Circos",
    "description": "Circos is a software package for visualizing data and information. It a
    "version": "0.69-4",
    "prerequisite": [],
    "input_list": [],
    "output_list": []
  },
  {
    "step_number": "8",
    "name": "#SBG_Bedpe4Oncofuse",
    "description": "SBG Bedpe4Oncofuse is a simple one-liner that prepares ChimeraScan BEDP
    "version": "1.0",
    "prerequisite": [],
    "input_list": [],
    "output_list": []
  },
  {
    "step_number": "9",
    "name": "#Circos_1",
    "description": "Circos is a software package for visualizing data and information. It a
    "version": "0.69-4",
    "prerequisite": [],
    "input_list": [],
    "output_list": []
  },
  {
    "step_number": "10",
    "name": "#SBG_Html2b64",
    "description": "Tool for converting HTML reports of FastQC, SnpEff, MultiQC (simple rep
    "version": "1.0",
    "prerequisite": [],
    "input_list": [],
```

```
      "output_list": []
    },
    {
      "step_number": "11",
      "name": "#SBG_Compressor",
      "description": "SBG Compressor performs the archiving(and/or compression) of the files
      "version": "v1.0",
      "prerequisite": [],
      "input_list": [],
      "output_list": []
    },
    {
      "step_number": "12",
      "name": "#SBG_Oncofuse4Circos",
      "description": "An R script that extracts fusion links and names needed from the Oncofu
      "version": "1.0",
      "prerequisite": [],
      "input_list": [],
      "output_list": []
    },
    {
      "step_number": "13",
      "name": "#ChimeraScan_Run",
      "description": "ChimeraScan detects fusion genes (chimeras) in paired-end RNA-seq datas
      "version": "0.4.5",
      "prerequisite": [],
      "input_list": [],
      "output_list": []
    }
  ]
},
"execution_domain": {
```

```
    "script": [
      "https://cgc-api.sbgenomics.com/v2/apps/phil_webster/bco-cwl-examples/fusion-transcript-
    ],
    "script_driver": "Seven Bridges Common Workflow Language Executor",
    "software_prerequisites": [],
    "external_data_endpoints": [],
    "environment_variables": []
  },
  "parametric_domain": [],
  "io_domain": {
    "input_subdomain": [
      {
        "uri": [
          {
            "filename": "",
            "uri": "",
            "access_time": ""
          }
        ]
      }
    ],
    "output_subdomain": [
      {
        "mediatype": "",
        "uri": [
          {
            "uri": "",
            "access_time": ""
          }
        ]
      }
    ]
```

```
  },
  "error_domain": {
    "empirical_error": [],
    "algorithmic_error": []
  }
}
```