DeviCNV

Version 1.0

1. Introduction

DeviCNV is a tool written in python (https://www.python.org/) and R (www.r-project.org) for detecting and visualizing exon-level copy number variants (CNVs) in targeted NGS data. DeviCNV has two submodules to be used in two popular targeted NGS approaches: hybridization capture- and polymerase chain reaction (PCR) capture-based approaches.

DeviCNV predicts a copy-number ratio with a confidence interval for each target capture probe (or amplicon) from bootstrapped regression models, and merges adjacent probe level CNV candidates together into a single larger CNV candidate. DeviCNV assigns a confidence level to every CNV candidate using a novel scoring scheme, and provides plots for visual inspection. Hereafter, we use the terms “probe” and “amplicon” interchangeably without losing generality with respect to read depth for the calculation of capture intervals.

2. Software environment

DeviCNV runs on Python 2.7 and R 3.2.0.

A. Python dependencies

i. sys, operator, random
ii. intervaltree, intervaltree_bio
iii. numpy
iv. scipy
v. pysam
vi. pyvcf

B. R dependencies

i. ggplot2
ii. PSCBS

3. Input

DeviCNV requires three inputs.

A. A set of analysis-ready binary alignment/map (BAM) formatted files

i. It is recommended to use the samples generated from the same batch of the sequencing experiment.

ii. It is recommended to use a minimum number (≥10) of input samples.

B. A txt file that contains the genomic position and pool information about target capture probes

i. This file must be TSV (tab-delimited file) format, and the nine columns are mandatory:

- Amplicon_ID (Amplicon/probe unique ID)
- Chr (Chromosome)
- Amplicon_Start (The amplicon/probe primer start genomic position, zero-based system)
- Insert_Start (For amplicon, the insert start genomic position, zero-based system; for bait, write "Amplicon_Start".)
- Insert_End (For amplicon, the insert end genomic position; for bait, write "Amplicon_End".)
- Amplicon_End (The amplicon/probe end genomic position)
- Gene (Single gene symbol where the probe is located.)
- Transcript (Transcript ID you used)
- Exon (Exons of the transcript where the probe is located.)
- Pool (The pool to which the target capture probe belongs; if there is no pool, write "Pool1".)

C. A txt file that contains sample's sex information

i. This file must be TSV (tab-delimited file) format, and the two columns are mandatory:

- Sample (Sample unique ID)
- Sex (Female or Male)

D. A txt file that contains threshold for DeviCNV's scoring system

i. This file must be TSV (tab-delimited file) format, and the three columns are mandatory:

- Attribute (The name of the measures to score)
- delFilter (A threshold for deletion)
- dupFilter (A threshold for duplication)

ii. We recommended our suggested default parameter setting.

Attribute	Default parameter setting	delFilter	dupFilter
ProbeCntInRegion	1 point for ≥ 2	>=2	>=2
AverageOfReadDepthRatios	If deletion, 1 point for < 0.65; If duplication, 1 point for > 1.35	<0.65	>1.35
STDOfReadDepthRatios	1 point for < 0.25	<0.25	<0.25
AverageOfCIs	1 point for < 0.3	<0.3	<0.3
AverageOfR2vals	1 point for ≥0.8	>=0.8	>=0.8

4. Running DeviCNV

DeviCNV consists of nine sequential Python and R scripts. An example script that runs DeviCNV can be found in GitHub with sample input files. (https://github.com/SD-Genomics/DeviCNV)

A. STEP 1 : Calculation of probe (or amplicon) level read depth ratio.

```
python python.calculateReadsDepthOfAmp.py <inSample> <inBamdir>
<inAmpliconTxt> <readDepthDir> <readDepthStatDir> <dedupOp> <datType>
<MQList>
```

Required arguments

<inSample>	A sample name of bam file
<inBamdir>	A directory that contains input bam files
<inAmpliconTxt>	A txt file that contains the genomic position and pool information about target capture probes
<readDepthDir>	A directory for a read depth per amplicon output file
<readDepthStatDir>	A directory for a read depth statistics output file
<dedupOp>	If true, use only reads after removing deduplicates; True is recommended.
<datType>	A NGS data type, HYB for hybridization-based approach and PCR for PCR-based approach.

<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. “MQ0,MQ10,MQ20”; "MQ0" is recommended.

B. STEP 2 : Generation single file that contains all input samples’ read depth statistics.

```
python python.mergeReadDepthStatistics.py <batchTag> <inSampleInfoTxt>
<readDepthStatDir>
```

Required arguments

<batchTag>	A prefix of output files
<inSampleInfoTxt>	A txt file that contains sample’s sex information
<readDepthStatDir>	A directory for a read depth statistics file

C. STEP 3 : X chromosome normalization of probe (or amplicon) level read depth ratio.

```
python python.chrXNormalizeReadDepth.py <batchTag> <inSampleInfoTxt>
<readDepthDir> <norReadDepthDir> <PoolList> <MQList>
```

Required arguments

<batchTag>	A prefix of output files
<inSampleInfoTxt>	A txt file that contains sample’s sex information
<readDepthDir>	A directory for a read depth per amplicon output file
<norReadDepthDir>	A directory for an output file that contain X chromosome normalized read depths
<PoolList>	A pool list of amplicon primers of your data
<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. “MQ0,MQ10,MQ20”; "MQ0" is recommended.

D. STEP 4 : Filtering low-quality samples.

```
Rscript r.filterLowQualSample.r <batchTag> <inSampleInfoTxt>
<norRDDir> <lqSampleDir> <MQList>
```

Required arguments

<batchTag>	A prefix of output files
<inSampleInfoTxt>	A txt file that contains sample’s sex information
<norReadDepthDir>	A directory for an output file that contain X chromosome normalized read depths
<lqSampleDir>	A directory for an output txt file and a pdf file that contain low-quality sample filtering results

<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. “MQ0,MQ10,MQ20”; "MQ0" is recommended.

E. STEP 5 : Building linear regression models with bootstrapping.

```
python python.buildLinearRegression.py <batchTag> <readDepthStatDir>
<norReadDepthDir> <lqSampleDir> <RCRatioDir> <MQList> <dupdelList>
```

Required arguments

<batchTag>	A prefix of output files
<readDepthStatDir>	A directory for a read depth statistics file
<norReadDepthDir>	A directory for a txt file that contain X chromosome normalized read depths
<lqSampleDir>	A directory for a txt file and a pdf file that contain low-quality sample filtering results
<RDRatioDir>	A directory for a out txt file that contains predicted read depth ratio by linear regression models
<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. “MQ0,MQ10,MQ20”
<dupdelList>	User-defined duplication/deletion thresholds for calculating duplication/deletion p-values; Ex. "1.2_0.8,1.3_0.7"; "1.2_0.8" is recommended.

F. STEP 6 : Generating two type of plot per sample, (1) a plot for whole gene and (2) the gene centric plots.

```
Rscript r.plotPerSample.r <batchTag> <inSample> <RDRatioDir> <plotDir>
<CBSDir> <PoolList> <MQList> <dupdelList>
```

Required arguments

<batchTag>	A prefix of output files
<inSample>	A sample name of bam file
<RDRatioDir>	A directory for a txt file that contains predicted read depth ratio by linear regression models
<plotDir>	A directory for output pdf files: two types of plots (1) a plot for whole gene and (2) the gene centric plots.
<CBSDir>	A directory for a output txt file that contains results of the circular binary segmentation (CBS) method to segment read depth profile.
<PoolList>	A pool list of amplicon primers of your data

<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. “MQ0,MQ10,MQ20”; "MQ0" is recommended.
<dupdelList>	User-defined duplication/deletion thresholds for calculating duplication/deletion p-values; Ex. "1.2_0.8,1.3_0.7"; "1.2_0.8" is recommended.

G. STEP 7 : Getting CNV candidates by the circular binary segmentation (CBS) approach and small region extraction approach.

```
python python.getCNVPerSample.py <batchTag> <inSample> <RDRatioDir>
<CBSDir> <CNVDir> <MQList> <dupdelList>
```

Required arguments

<batchTag>	A prefix of output files
<inSample>	A sample name of bam file
<RDRatioDir>	A directory for a txt file that contains predicted read depth ratio by linear regression models
<CBSDir>	A directory for a txt file that contains results of the circular binary segmentation (CBS) method to segment read depth profile.
<CNVDir>	A directory for a output txt file that contains CNV candidates
<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. “MQ0,MQ10,MQ20”; "MQ0" is recommended.
<dupdelList>	User-defined duplication/deletion thresholds for calculating duplication/deletion p-values; Ex. "1.2_0.8,1.3_0.7"; "1.2_0.8" is recommended.

H. STEP 8 : Merging samples’ CNV candidate files into the single txt file.

```
python python.mergeCNVFiles.py <batchTag> <inSampleInfoTxt> <CNVDir>
<MQList> <dupdelList>
```

Required arguments

<batchTag>	A prefix of output files
<inSampleInfoTxt>	A txt file that contains sample’s sex information
<CNVDir>	A directory for txt files that contains CNV candidates
<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. “MQ0,MQ10,MQ20”; "MQ0" is recommended.

<dupdelList>	User-defined duplication/deletion thresholds for calculating duplication/deletion p-values; Ex. "1.2_0.8,1.3_0.7"; "1.2_0.8" is recommended.

I. STEP 9 : Scoring all CNV candidates by using DeviCNV's unique scoring system.

```
python python.scoreCNV.py <batchTag> <CNVDir> <lqSampleDir> <MQList>
<dupdelList> <inScoringThTxt>
```

Required arguments

<batchTag>	A prefix of output files
<CNVDir>	A directory for a txt file that contains CNV candidates
<lqSampleDir>	A directory for a txt file and a pdf file that contain low-quality sample filtering results
<MQList>	Mapping quality value (MQV) thresholds for using reads with more than the MQV; Ex. "MQ0,MQ10,MQ20"; "MQ0" is recommended.
<dupdelList>	User-defined duplication/deletion thresholds for calculating duplication/deletion p-values; Ex. "1.2_0.8,1.3_0.7"; "1.2_0.8" is recommended.
<inScoringThTxt>	A txt file that contains threshold for DeviCNV's scoring system

5. Output

A. SAMPLE.readCount.POOL.txt

i. This output file contains the read depth per amplicon including the pool for the sample.

Amplicon_ID	Amplicon/probe unique ID
Chr	Chromosome
Amplicon_Start	The amplicon/probe primer start genomic position, zero-based system
Insert_Start	For amplicon, the insert start genomic position, zero-based system; for bait, write "Amplicon_Start".
Insert_End	For amplicon, the insert end genomic position; for bait, write "Amplicon_End".
Amplicon_End	The amplicon/probe end genomic position
Gene	Single gene symbol where the probe is located.
Transcript	Transcript ID you used
Exon	Exons of the transcript where the probe is located.
Pool	The pool to which the target capture probe belongs.
MQs	Read depths according to the mapping quality threshold.

B. SAMPLE.readDepthStatistics.txt

i. This output file contains the read depth statistics of the sample.

Sample	SAMPLE
Pool	POOL
MQ	Mapping quality threshold
Mean	The mean of read depths of all amplicons including the pool
Median	The median of read depths of all amplicons including the pool
StandardDeviation	The standard deviation of read depths of all amplicons including the pool
Sum	The sum of read depths of all amplicons including the pool

C. PREFIX.All.readDepthStatistics.txt

i. This output file contains all samples' read depth statistics.

D. PREFIX.readDepth.normalizedChrX.MQTH.txt

i. This output file contains all samples' read depth after X chromosome normalization.

Amplicon_ID	Amplicon/probe unique ID
Chr	Chromosome
Amplicon_Start	The amplicon/probe primer start genomic position
Insert_Start	For amplicon, the insert start genomic position
Insert_End	For amplicon, the insert end genomic position
Amplicon_End	The amplicon/probe end genomic position
Gene	Single gene symbol where the probe is located.
Transcript	Transcript ID you used
Exon	Exons of the transcript where the probe is located.
Pool	The pool to which the target capture probe belongs.
SAMPLEs	Normalizaed read depths of the amplicon of the sample

E. PREFIX.All.lowQualitySampleTest.MQTH.txt

i. DeviCNV filter out a sample that had a low coefficient of correlation with other samples in the same input set. (By default, 75 percentile of correlation coefficients < 0.7)

Sample	SAMPLE
0%	The minimum value of correlation coefficients
25%	The 25 percentile value of correlation coefficients
50%	The median value of correlation coefficients

75%	The 75 percentile value of correlation coefficients
100%	The maximum value of correlation coefficients
LowQualSample?	If the sample's 75 percentile value is lower than, 0.7, then "LQ".

F. PREFIX.All.lowQualitySampleTest.MQTH.pdf

i. This plot shows the samples' read depth correlation coefficient between other samples.

ii. The Dashed line shows low-quality sample filter threshold, 0.7.

G. PREFIX.readDepthRatioFromLRModel.MQTH.DUPDELTH.txt

i. This output file contains predicted read depth ratio from linear regression models.

Amplicon_ID	Amplicon/probe unique ID
Chr	Chromosome
Amplicon_Start	The amplicon/probe primer start genomic position, zero-based system
Insert_Start	For amplicon, the insert start genomic position, zero-based system; for bait, write "Amplicon_Start".
Insert_End	For amplicon, the insert end genomic position; for bait, write "Amplicon_End".
Amplicon_End	The amplicon/probe end genomic position
Gene	Single gene symbol where the probe is located.
Transcript	Transcript ID you used
Exon	Exons of the transcript where the probe is located.
Pool	The pool to which the target capture probe belongs.
MQ	MQTH
Type	MedianRD, the median of the SAMPLE read depth of the POOL; Y, the observed read depth; Y_L, the low bound of the predicted read depth; Y_M, the median of predicted read depth; Y_U, the upper bound of the predicted read depth; CI_L, the low bound of 95 % confidence interval of predicted read depth ratio; CN_M, the median of 95% confidence interval of predicted read depth ratio (=copy number, CN); DupPval, the p-value for duplication; DelPval, the p-value for deletion; Pvalue, the smaller of DupPval and DelPval; CNVType, duplication/deletion/neutral/faultySample/faultyAmp; regRval, the mean of R-squared values.
SAMPLEs	the value of the type of the SAMPLE

H. MQTH.DUPDELTH _SAMPLE _AllGene.pdf

i. This output file shows the overall result for one sample across whole genes.

1. The title : the sample name and the median of the sample read depths for each pool.
2. Gray dotted lines : duplication/deletion thresholds
3. The point shape : pool information and faulty/low-quality types
4. The red gradient : the p-value

I. MQTH.DUPDELTH _SAMPLE_GENE.pdf

i. This output file shows the gene centric plot containing detailed information.

1. The title : the sample name and the median of the sample read depths for each pool.
2. Gray dotted lines : duplication/deletion thresholds
3. The point shape : pool information and faulty/low-quality types
4. The red gradient : the p-value
5. The bar of the point : the 95% confidence interval of the predicted read depth ratio.
6. Grey boxes : Exon information where each amplicon is located.

J. SAMPLE.CBS.MQTH.DUPDELTH.tsv

i. This output file contains the results of the CBS of the SAMPLE.

K. SAMPLE.CNV.MQTH.DUPDELTH.txt

i. This output file contains CNV candidates of the SAMPLE.

L. PREFIX.All.CNV.MQTH.DUPDELTH.txt

i. This output file contains all samples' CNV candidates.

M. PREFIX.All.CNVWithScore.MQTH.DUPDELTH.txt

i. This output file contains CNV candidates for all samples with scores and related information.

Score	The score of this CNV candidate
RegionID	The ID of this CNV candidate

CnvType	The Type of this CNV candidate
Sample	The sample name
MedianRDOfSample	The median of read depth of the sample
Chr	Chromosome
Start	The start genomic position of this CNV candidate
End	The end genomic position of this CNV candidate
Gene	The gene where this CNV candidate is located
Transcript	The transcript ID of the gene
ExonCntInRegion	The number of exons covered by this CNV candidate
CoveredExonCnt	The number of exons covered by the user-designed amplicon within this gene
ExonInRegionRatio	ExonCntInRegion/CoveredExonCnt
Exons	The list of exons belonging to this CNV candidate
AmpCntInRegion	The number of amplicons in this CNV candidate
TotalAmpCnt	The number of amplicons in this gene, except amplicons with faulty type.
Amplicons	The list of amplicons in this CNV candidate
Pvalues	The list of p-values in this CNV candidate
FilterInAmpRatio	The ratio of amplicons in this CNV candidate that has a p-value of less than 0.05.
AverageOfReadDepthRatios	The average of predicted read depth ratio in this CNV candidate
STDOfReadDepthRatios	The standard deviation of predicted read depth ratio in this CNV candidate
ReadDepthRatios	The list of read depth ratios in this CNV candidate
AverageOfCIs	The average of confidence interval lengths of predicted read depth ratios in this CNV candidate
AverageOfR2vals	The average of R-squared values of linear regression model for predicting read depth ratios in this CNV candidate

6. License

DeviCNV is under GNU license (GPLv3).