18 September 2018

SNATCNV: Single Nucleotide Association Test for CNV analysis

Page: https://github.com/hamidrokny/SNATCNV

1) What is SNATCNV?

SNATCNV is a powerful CNV association toolset to robustly identify recurrently deleted and duplicated regions from microarray, whole exome, and whole genome sequencing that are significantly associated with disease/disorder (e.g. autism). For every position in the genome, SNATCNV counts how often the nucleotide is either deleted or duplicated in the case and control CNVs. It then uses the one-tailed Fisher's exact test to determine whether the base is significantly more frequently duplicated or deleted in cases or controls. To identify significant regions, SNATCNV calculated *P*-values for 500,000 random permutations of case/control labels to estimate the probability that an association emerges by chance.

---

2) How to install it?

The SNATCNV pipeline requires the following dependencies:

2.1) MATLAB version

MATLAB software (>2012)

2.2) R version

R (>2.2.1)

RColorBrewer and ggplot2 (>2.2.1) packages

---

3) Input files

In order to process the CNV dataset, SNATCNV requires two files as following:

3.1) CNV dataset

SNATCNV takes CNV dataset in the following format:

chr: Chromosome information of a CNV.

chr: Chromosome information of a CNV (should be numeric. Put 23 for X, 24 for Y and 25 for M).

start: Starting position of a CNV.

end: Ending position of a CNV.

PatientID: Sample ID.

CNV_type: should be “Del” or “Dup”.

case_control: should be “case” or “control”.

Patient_gender: Can be “M” or “F”; if you have no gender information, put ”NA”.

Table.1 Example of the CNV dataset.

chr	chr	start	end	patientID	CNV_type	case/control	patient_gender
chr1	1	10001	10077413	A12764	Del	case	M
chr1	1	147198154	148498814	A12766	Dup	case	F
chr2	2	51237623	52359612	A12766	Del	control	F
chr2	2	234091785	239251404	A87021	Del	control	M
chr6	6	7601267	7701391	A61230	Dup	NA	NA
chr6	6	2045813	2145514	A61346	Del	case	NA
chr11	11	19435001	19436039	A12317	Del	case	F
chrX	23	26534242	27538142	A98125	Dup	control	M

3.2) List of protein-coding and non-coding genes

SNATCNV accepts gene annotations in the following format.

Table.2 Example of the gene lists as an input for SNATNV.

GeneID	**Gene_symbol**	**Gene_class**	**Gene_chr**	**Gene_start**	**Gene_end**	**Strand**
ENSG00000225880.4	LINC00115	lncRNA_sense_intronic	chr1	91421	762886	-
ENSG00000240453.1	RP11-206L10.10	lncRNA_intergenic	chr1	91421	762886	-
CATG00000042732.1	CATG00000042732.1	lncRNA_divergent	chr1	523500	569943	-
ENSG00000181790.6	BAI1	coding_mRNA	chr8	143530811	143626370	+
CATG00000104410.1	CATG00000104410.1	coding_mRNA	chr8	143548473	143557002	-
CATG00000104412.1	CATG00000104412.1	short_ncRNA	chr8	143647895	143648677	-
CATG00000101268.1	CATG00000101268.1	coding_mRNA	chr8	143660416	143688117	+
CATG00000022111.1	CATG00000022111.1	lncRNA_divergent	chr14	102782989	102785943	-
ENSG00000022976.11	ZNF839	coding_mRNA	chr14	102783528	102815203	+
CATG00000113577.1	CATG00000113577.1	lncRNA_intergenic	chrX	42540546	42615236	-
ENSG00000102055.5	PPP1R2P9	pseudogene	chrX	42636722	42637412	-
ENSG00000224035.1	SFPQP1	pseudogene	chrY	15195728	15207719	+
CATG00000115082.1	CATG00000115082.1	lncRNA_intergenic	chrY	15326352	15336018	-
ENSG00000183878.11	UTY	coding_mRNA	chrY	15345781	15591867	-

Note: strand can be “NA”.

---

4) How to use SNATCNV?

4.1) Script "SNATCNV_indvbased"

4.1.1) Setting arguments

The following parameters should be set to run this script:

SIGNIFICANT_TRD=0.006.

Permutation cut-off P value to identify significant CNV regions. This cut-off may be different for deletion and duplication regions.

CNV_TYPE='Del'.

Can be specify by 'Del', 'Dup' or 'Both'.

CNV_SIZE=1000.

Minimum length (base pair) for a CNV; all CNVs with less than this size will be excluded from the analysis.

GENOME_BUILD=GENOME_BUILD_19.

Can be specify by two terms 'GENOME_BUILD_18' or 'GENOME_BUILD_19'.

GENELIST='FANTOM_CAT.lv3_robust.genes.tsv'.

Path to ENCODE or FANTOM5 gene list.

CNV_FILENAME='AutDB_ASD_cnv_dataset.txt'.

Path to CNV dataset.

OUTPUT_SIGNIFICANT_REGIONS_ALL='significant_regions_based_on_indv_ALL.txt'.

Path to save full list of significant regions identified by SNATCNV.

OUTPUT_SIGNIFICANT_REGIONS_SUMMARY='significant_regions_based_on_indv_SUMMARY.txt'.

Path to save a summary list of significant regions identified by SNATCNV.

OUTPUT_SIGNIFICANT_GENES='significant_genes_based_on_indv.txt'.

Path to save significant genes within significant regions identified by SNATCNV.

4.1.2) Output files

SNATCNV_indvbased script generates outputs in different formats as following:

a) A complete map of CNV regions. Table 3 shows an example of this output.

Table3. An overview of SNATCNV output as a complete map of identified CNV regions.

chr	start	end	number_of_case	number_of_control	Pvalue_case	Pvalue_control
1	1	10000	1	0	7.41E-01	1.00E+00
1	10001	61722	79	0	5.00E-11	1.00E+00
1	61723	98590	80	0	3.70E-11	1.00E+00
15	30073735	30092905	24	1	5.43E-03	9.99E-01
15	30092906	30205914	22	0	1.37E-03	1.00E+00
23	7323930	7381307	71	10	2.32E-03	9.99E-01
23	7381308	7401203	71	11	4.75E-03	9.98E-01
23	155250659	155251871	16	0	8.28E-03	1.00E+00
23	155251872	155270560	1	0	7.41E-01	1.00E+00
24	1	10678	0	0	1.00E+00	1.00E+00
24	10679	175736	3	0	4.07E-01	1.00E+00
24	59337897	59354877	5	0	2.24E-01	1.00E+00
24	59354878	59373566	0	0	1.00E+00	1.00E+00

b) A full list of significant CNV regions based on the permutation cut-off. Table 4 shows an example of this output.

Table4. An overview of SNATCNV output as a full list of significant CNV regions (permutation cut-off is 0.006).

chr	start	end	number_of_case	number_of_control	Pvalue_case	Pvalue_control
1	10001	61722	79	0	5.00E-11	1.00E+00
1	61723	98590	80	0	3.70E-11	1.00E+00
15	30073735	30092905	24	1	5.43E-03	9.99E-01
15	30092906	30205914	22	0	1.37E-03	1.00E+00
23	7323930	7381307	71	10	2.32E-03	9.99E-01
23	7381308	7401203	71	11	4.75E-03	9.98E-01

c) A summary of significant regions based on the permutation cut-off.

Here, SNATCNV merges the continues small regions to make a bigger significant region with Pvalue less than the threshold. For example, two regions chr1:10001-61722 and chr1:61723-98590 will be merged and make a bigger region chr1:10001-98590. Table 5 shows an example of this output.

Table5. An overview of SNATCNV output as a summary of significant regions based on the permutation cut-off (permutation cut-off is 0.006).

chr	start	end	number_of_case	number_of_control	Pvalue_case	Pvalue_control
1	10001	98590	80	0	3.70E-11	1.00E+00
15	30073735	30205914	24	1	5.43E-03	9.99E-01
23	7323930	7401203	71	11	4.75E-03	9.99E-01

d) List of all coding and non-coding genes within the significant regions. Table 6 shows an example of this output.

Table6. An overview of SNATCNV output for significant genes (permutation cut-off is 0.006).

GeneID	Gene_symbol	Gene_class	Gene_chr	Gene_start	Gene_end	Strand	CNV_type
ENSG00000225880.4	LINC00115	lncRNA_sense_intronic	1	91,421	762,886	-	Del
ENSG00000240453.1	RP11-206L10.10	lncRNA_intergenic	1	91,421	762,886	-	Del
CATG00000042732.1	CATG00000042732.1	lncRNA_divergent	1	523,500	569,943	-	Del
ENSG00000223659.1	RP5-857K21.5	lncRNA_divergent	1	562,757	564,475	-	Del
CATG00000042733.1	CATG00000042733.1	lncRNA_divergent	1	567,299	567,595	-	Del
ENSG00000228327.2	RP11-206L10.2	lncRNA_sense_intronic	1	676,386	762,886	-	Del
ENSG00000237491.4	RP11-206L10.9	lncRNA_intergenic	1	714,172	740,255	+	Del
ENSG00000177757.1	FAM87B	lncRNA_intergenic	1	753,349	755,214	+	Del
ENSG00000228794.4	RP11-206L10.11	lncRNA_intergenic	1	762,215	809,715	+	Del
ENSG00000269308.1	AL645608.2	coding_mRNA	1	817,585	819,983	+	Del

---

4.2) Script "SNATCNV_permutation"

4.2.1) Setting arguments

CNV_TYPE='Del'.

Can be specified as 'Del', 'Dup' or 'Both'.

CNV_SIZE=1000.

Minimum length (base pair) for a CNV; all CNVs with less than this size will be excluded from the analysis.

CNV_FILENAME='AutDB_ASD_cnv_dataset.txt'.

Path to CNV dataset.

OUTPUT_PERMUTATION= 'permutation_ASD_del.txt'.

Path to save permutation output.

MAX_CNV=600.

Maximum number of case and control samples with CNV that overlap with the same position.

NUMBER_OF_ITERATION=500000.

Number of permutation.

4.2.2) Output files

Table7. An overview of SNATCNV permutation output

confidence interval	Pvalue_threshold	CNV_type
C.I=0.9999	2.50E-08	Del

C.I=0.999	5.00E-08	Del
C.I=0.99	8.50E-08	Del
C.I=0.98	3.20E-07	Del
C.I=0.97	9.70E-06	Del
C.I=0.96	2.10E-05	Del
C.I=0.95	3.20E-05	Del
C.I=0.94	9.20E-05	Del
C.I=0.93	4.90E-04	Del
C.I=0.92	6.50E-04	Del
C.I=0.91	1.30E-04	Del
C.I=0.9	5.30E-03	Del
C.I=0.8	1.00E-02	Del
C.I=0.7	5.00E-02	Del
C.I=0.6	2.00E-01	Del
C.I=0.5	5.60E-01	Del

---

5) Required programs

SNATCNV is available for MATLAB and R users.

MATLAB: https://www.mathworks.com

R: https://www.r-project.org

---

6) Test Dataset

The AutDB CNV dataset (http://autism.mindspec.org/autdb); download date: July 2016, are available in the SNATCNV github page (https://github.com/hamidrokny/SNATCNV).

---

7) Contact

Hamid Alinejad-Rokny <h(dot)alinejad(at)ieee(dot)org>, Centre for Medical Research, The University of Western Australia, Perth, 6008, Australia.