

# AB CNV Tool Documentation

<b>1</b>	<b>OVERVIEW .....</b>	<b>2</b>
<b>2</b>	<b>INSTALLATION .....</b>	<b>3</b>
2.1	Prerequisites .....	3
2.2	Installation procedure .....	3
<b>3</b>	<b>AB CNV TOOL.....</b>	<b>4</b>
3.1	Algorithm/Script Description.....	4
3.2	Usage Parameters – Required.....	6
3.3	Usage Parameters - Optional .....	7
3.4	Usage Example .....	8
3.5	List of programs/Scripts Included .....	8
3.6	Other Scripts Called by this Program .....	8
3.7	Path Constraints.....	8
3.8	System Input Files.....	8
3.9	Input File Versions Supported .....	8
3.10	Output Files.....	9

1.0	7/10/2009	
-----	-----------	--

## 1 Overview

The AB CNV Tool detects copy number variation using SOLiD™ system data from a single human sample mapped to the human reference sequence hg18. The program detects copy number variations (CNVs) in a SOLiD™ system run of a single human sample, without the need for a matched normal sample.

The program takes, as input, the SOLiD™ GFFv2 files (or .mates files) containing uniquely mapped reads. From these files, the program calculates coverage and computes the log ratios from the data and normalizes the log ratios by predicted mappability and by GC content. The raw/normalized log ratios are then smoothed and segmented using a Hidden Markov Model to generate CNV Calls. Windows with CNV calls are merged into segments and user-defined set of filtering criteria are applied on the segments to exclude low probability CNV segments.

The program also requires predicted mappability files and reference sequence files (.fa), which are provided. For this version, these pre-computed files are available only for the human reference sequence hg18; so the first version of the program only supports human CNV detection.

By default, the pipeline does **not** call CNVs within 1 MBase of the centromeres and telomeres of the chromosomes. These are highly repetitive regions and tend to contain a lot of apparent CNVs that may not be of biological interest. The distance from the centromeres and telomeres in which CNVs are not called is a parameter that you can set to any desired distance.

## 2 Installation

### 2.1 Prerequisites

Before installation, verify that your system meets the following requirements:

- Linux CentOS 4 – The program has been built and tested on Linux CentOS 4 using GCC 3.4 and GNU Make 3.8.
- The pipeline requires up to 8 GB of RAM to run, depending upon the window size used for sampling.
  - For default window size of 5000 bases, it requires 2GB of RAM
  - For window size of 2000 bases, it requires 4GB of RAM
  - For window size  $\leq 1000$  bases, it requires  $\geq 8$ GB of RAM

### 2.2 Installation procedure

To install the program, follow these steps:

1. Unzip the contents of the `SOLiD_CNV_<version>.tar.gz` package into the desired location.
2. Enter the subdirectory created and run `make`.

This command will build the tool from source, creating the main executable `SOLiD_CNV_TOOL`. To test the integrity of the distribution, run `make test`. The software should run to completion with no errors.

3. Install the binary to a location in the PATH if desired.

Use the `make install` target and set the `PREFIX` variable to the installation directory root. For example, to install into an existing Corona Lite package, enter the following:

```
$ make install PREFIX=$CORONAROOT
```

This target simply copies the executable to a subdirectory called `bin` underneath the `PREFIX` root.

To install the supporting data (mappability files and reference sequences), follow these steps:

1. Unzip the contents of the `hs_CNV_data_<version>.tar.gz` package into the desired location for the reference data files. These files are large and reference-specific and so are less likely to change between revisions of the software. It is reasonable to place these files along side other genomic reference data.
2. Update the “`data.dir`” value in the `referenceMapping.cmap` file provided in the package to the path of the data directory containing mappability files and fasta files.

### 3 *AB CNV Tool*

**Program Name:** AB\_CNV\_TOOL

**Program Version:** 1.0

**Development Languages:** C (gcc compiler)

**Compiled for:** UNIX/LINUX

**Supports AB kit or protocol or sample prep method:** N/A

**PBS Required:** No

**Date:** June 3, 2009

#### 3.1 **Algorithm/Script Description**

The pipeline is comprised of these main algorithmic steps:

**Pre-Processing:** The coverage files are converted from input format (GFF or .mates) to binary format. User-defined configuration parameters are initialized. A working directory for intermediate files is created.

**Sampling & Normalization:** This step divides the chromosomal region into windows of variable size. The window size depends the mappability of the region, so that approximately the same number of mappable positions are in each window. The program computes the log ratios between coverage mean of every window and mean of the whole chromosome arm. It normalizes the log ratios based on GC Content using a run-specific normalization.

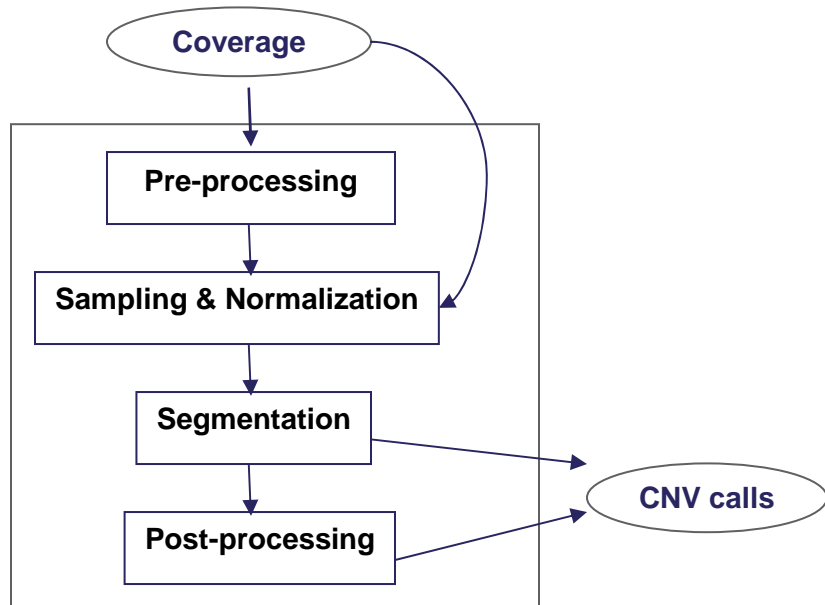
**Segmentation:** This step segments the log ratios data using a Hidden Markov Model (HMM). The HMM converts continuous log ratio values of the windows into discrete copy number states.

#### **Post-Processing:**

Neighboring windows are merged into a segment in the following cases:

1. When they have same copy number.
2. When they all have copy number deletions: Copy numbers less than 2 and copy numbers differ by 1.
3. When they all have similar copy number amplifications: Copy numbers greater than 2 and copy numbers differ by 1.

The program then applies various user-configurable filtering criteria on the CNV calls, such as Minimum mappability, number of continuous blocks, and so on.



### 3.2 Usage Parameters – Required

**Note:** The parameter order is **not** important.

Parameter	Description	Example	Acceptable Values
--exp-name	Name of the experiment. Typically contains the name of the sample, MP or Frag for mate-pair data and window size used for sampling.	NA19240_MP_cn v_5_5k	<ul style="list-style-type: none"> <li>All alpha numeric characters.</li> <li>Special characters allowed: Only – and _</li> </ul>
--data-type	MP if it is mate-pair. FRAG if it is fragment data. (Case-Sensitive)	MP	<ul style="list-style-type: none"> <li>MP</li> <li>FRAG</li> </ul>
--tag-length	Maximum read length.	25	Integer > 0
--cmap-file	Fully qualified path to the cmap file.	/home/cnvowner/cnv_data/referenceMapping.cmap	Absolute file path
--coverage-format	Format of coverage files provided. (Case-Sensitive)	GFF	<ul style="list-style-type: none"> <li>GFF</li> <li>Mates</li> <li>Text</li> <li>Binary</li> </ul>
--coverage-file	Input coverage files: <ul style="list-style-type: none"> <li>Path to the directory if Text or Binary</li> <li>Path to the file if GFF or Mates</li> </ul>	/home/cnvowner/data/coverageFiles	Absolute file path

### 3.3 Usage Parameters - Optional

Parameter	Description	Example	Acceptable Values
--window-size	Size of the window block to be considered as one region. Recommended sizes – 5000, 2000, 1000	5000	Integer > 0 Default: 5000
--trim-distance	Distance in Kilobases to be trimmed from the extreme ends of the chromosome arms.	1000	Integer > 0 Default: 1000
--output-dir	Fully qualified path to the output directory.	/home/cnvowner/output_cnv	Absolute file path Default: "output"
--max-pval	Maximum p-value of a CNV segment.	0.02	0 < pVal < 1 Default: 0.25
--uminmap	Minimum mappability percentage for the regions to be shown as having copy number < 2.	60.0	Float value > 0.0 and < 100.0 Default: 10.0
--ominmap	Minimum mappability percentage for the regions to be shown as having copy number < 2.	10.0	Float value > 0.0 and < 100.0 Default: 10.0
--uminblocks	Minimum number of continuous blocks with copy number < 2.	2	Integer < 0 Default: 2
--ominblocks	Minimum number of continuous blocks with copy number > 2.	2	Integer > 0 Default: 2
--max-log-ratio	Maximum log ratio threshold for copy number deletion regions. (Log ratio of CN=2 is 0.0, CN=1.25 is -0.678 and CN=1 is -1.0)	-0.678	Float value < 0.0 & > -1.0 Default: -0.678
--min-log-ratio	Minimum log ratio threshold for copy number amplification regions. (Log ratio of CN=2 is 0.0, and CN=2.6 is 0.378)	0.378	Float value > 0.0 & < 0.5 Default: 0.375

### 3.4 Usage Example

Use the directory `test_Env` in `hs_CNV_<version>` as a test environment to test the installation.

Install the CNV Pipeline (see above) and run `make test` or enter the `test_Env` directory and enter the following command:

```
../SOLiD_CNV_TOOL --exp-name Test_Run --data-type MP --tag-length 25
--coverage-format Binary
--coverage-file /path/to/cnv/dist/test_Env/inputCoverageFiles
--cmap-file /path/to/cnv/dist/test_Env/referenceMapping.cmap
--output-dir /path/to/cnv/dist/test_Env/Test_Run
```

Where `/path/to/cnv/dist` is the absolute path to the unpacked SOLiD™ Human Copy Number Variation source code.

This command runs the CNV analysis on "chromosome 11 of T8 patient data". A new directory, `Test_Run`, is created with all the intermediate files and output files that are generated. The folder `Results` in the `test_Env` directory contains these output files.

At any time, if there is no change in the code, all the files generated in the `Test_Run` directory should be exactly same as those files in the `Results` directory.

### 3.5 List of programs/Scripts Included

The main program in this pipeline is `SOLiD_CNV_Tool`.

### 3.6 Other Scripts Called by this Program

None.

### 3.7 Path Constraints

All file names given as input should include full paths, not relative paths.

### 3.8 System Input Files

The configuration file specifies all files used by the pipeline.

- The pipeline accepts only one Mates or GFF file as input. To analyze multiple slides together, all of the Mates files or GFF files should be concatenated (joined) to a single Mates, or GFF, file and that file should be used as input. The file does not need to be sorted.
- To speed up re-running of CNV detection with different filtering criteria, the program also accepts Text or Binary Coverage files from a previous analysis run instead of GFFv2 or .mates files.
- The cmap file, specifies all the reference files used by the pipeline. The cmap file is a required input.

### 3.9 Input File Versions Supported

Mates – Non redundant

GFF – V2 Unique reads only

### 3.10 Output Files

All the files generated are placed in the output directory `output.dir`.

- `output.out`: Contains the `cnv` calls across all the chromosome arms.
- `output_filtered.out`: Final output of `cnv` calls after applying filter conditions.
- `output_filtered.gff`: Final output of `cnv` calls in GFF format. The format is validated by GFF3 validator, and you can view the file in the UCSC Genome Viewer.

For every chromosome arm `chrXX`, the following intermediate files are generated during the execution of the pipeline:

- `combined_coverage_chrXX.txt`: Coverage files in Text format.
- `coverage_chrXX.bcv`: Coverage files in binary format.
- `chrXX_EXPNAME_yGC.lgr`: GC normalized log ratios file.
- `chrXX_EXPNAME_yGC.calls`: CNV calls after segmentation.
- `SOLID_CNV_LOG.TXT`: A log file generated in the output directory

### 3.10.1 Output Fields of “output.out” and “output\_filtered.out”

Column Name	Description	Example
Chrom	Chromosome Number.	Chr11
start	Start location of the CNV Region.	12762605
end	End location of the CNV Region.	12916184
mappability	Fraction of Mappable bases in the CNV Region.	91.492867
log2Ratio	Mean of Log2Ratios of the windows in the CNV Region.	-0.888121
copynumber	Copy number of the region.	1
numWindows	Number of windows in the CNV region.	28
p-val	p-Value of the CNV Call for the region.	0.006965
frAcceptability	Fraction of windows in the Region that passed all the filtering criteria individually.  (Filtering criteria includes minimum Mappability, minimum number of windows, min log ratio, max log ratio and max p-value)	82.142857

### 3.10.2 Output Fields of “output\_filtered.gff”

seqid	The ID of the sequence to which the start and end coordinates refer.	Chr11
source	Free text qualifier indicating the algorithm or method that generated the feature.	AB_CNV_PIPELINE
type	Sequence ontology derived type for this variation.	repeat_region
start	Start position of the CNV Region.	12939004
end	End position of the CNV Region.	12939004
score	p-Value of the CNV Region.	0.004516
Strand	Not used for this output.	
phase	Not used for this output.	
Attributes	Various attributes of the CNV Region including copynumber, log2Ratio, numWindows and mappability. Refer 3.10.1 for the description of each of these attributes.	copynumber=1;log2Ratio=-0.888121;numWindows=28;mappability=91.492867

### **Licensing**

This software is being licensed to you under the OSI compliant GNU Public License (GPL V3). The license can be found at the following URL: <http://www.gnu.org/licenses/gpl.html>. Please read the license in its entirety and ensure that you understand the licensing conditions for use. Your use of this software indicates your acceptance of this licensing agreement.

© 2009 Life Technologies Corporation. All rights reserved.

The trademarks mentioned herein are the property of Life Technologies Corporation or their respective owners.