

AB Large InDel Tool Documentation

1	OVERVIEW	2
2	INSTALLATION	2
2.1	Prerequisites	2
2.2	Installation procedure	2
3	AB LARGE INDEL TOOL	3
3.1	Algorithm/Script Description.....	3
3.2	Usage Parameters – Required.....	4
3.3	Usage Parameters – Optional.....	5
3.4	Usage Example	5
3.5	List of programs/Scripts Included	6
3.6	Other Scripts Called by this Program	6
3.7	Path Constraints.....	6
3.8	System Input Files	6
3.9	Input File Versions Supported	6
3.10	Output Files.....	7
3.11	Sample Input Files.....	8
3.12	Sample Output Files	8

1.0	07/10/2009	
-----	------------	--

1 Overview

This software identifies deviations in clone insert size that indicate intra-chromosomal structural variations compared to a reference genome.

Insertions and deletions (indels) up to 100Kb are inferred by identifying positions in the genome in which the pairing distance between mapped mate-pairs is significantly deviated from what is expected at the given level of clone coverage.

A look-up table is created in which the amount that the clones must be deviated to achieve one standard deviation of significance is the standard error at each level of clone coverage. This produces an asymptotic curve in which the minimum size of detectable indels at a given level of significance drops rapidly as the clone coverage increases. The look-up table is used to determine the significance of the deviation in average insert size at each position in the genome. Regions of the genome that are significantly deviated are selected as candidate indels and hierarchical clustering is used to segregate the clones into groups in which the difference in the sizes of all clones in a group is less than the range specified by the user.

Clusters with too few clones (as specified by the user) are removed and the candidates are assessed to see if there remains a homozygous or heterozygous population of deviated insert sizes.

All clones deviated by ≥ 100 kb are discarded. Clones from various libraries with various insert sizes contribute to a single indel call by combining the probabilities associated with the clones from each library.

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 Perl 5.8.5 and GNU Make 3.8.
- This pipeline requires an existing Corona Lite (v4.2) installation to run.

2.2 Installation procedure

To install the standalone package into an existing Corona Lite installation, follow these steps:

1. Unzip the contents of the `Large_Indel_<version>.tar.gz` package into the desired location.
2. Enter the subdirectory created and enter the following:

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

where `$CORONAROOT` is the root of the Corona Lite installation. This will install the appropriate executable scripts into the `bin` directory.

- As with the Corona Lite tools, ensure that the `CORONAROOT` variable is set to the root of the installation and source of the `$(CORONAROOT)/etc/profile.d/corona.sh` file. The Large InDel tool should be ready to run.

3 AB Large InDel Tool

Program Name: Large InDel Tool

Program Version: 1.0

Development Languages: Perl

Supports AB kit or protocol or sample prep method: This software **only** supports protocols in the SOLiD™ system library guides (does not support HPLC size selection).

PBS Required: No

Date: June 19, 2009

3.1 Algorithm/Script Description

Pre-processing

During the upstream pairing and mate-pair rescue, pipelines clones are pre-classified as AAA, AAB, or AAC depending on the orientation and length distributions of individual mate-pairs. (See the table below for a review of these categories or README_PAIRING in the release notes for a more detailed description). Pre-processing involves removing discrepant-length clones with predicted insert sizes > 100kb. Putative inversions, such as BA*, and other chromosomal abnormalities, such as AB*, are also eliminated.

Category ^a	A Reference	B Insertion	C Deletion
AA = Same strand and orientation.	AAA	AAB	AAC
AB = Same strand, reverse orientation.	ABA	ABB	ABC
BA = Opposite strands, reads oriented away from each other.	BAA	BAB	BAC
BB = Opposite strands, reads oriented towards each other.	BBA	BBB	BBC

^aC** = reads map to different sequences

Calculating inter-read distances and generating the look-up table

The look-up table maps individual levels of clone coverage to a sample-corrected variability measurement, based on the population of mate-paired clones with the expected distance, order and orientation, such as AAA clones.

The standard deviation (SD) of the sizes of these clones is calculated and a look-up table is generated by calculating the standard error (SE) for each level of clone coverage, up to a maximum defined by the user (using the `-max-clone-cov` option). This produces an asymptotic curve where the minimum detectable indel size drops

rapidly as clone coverage increases. Multiple libraries are processed by calculating a unique look-up table for each mates file.

Calculating clone coverage and average insert size

For each genomic location, the algorithm calculates the number of spanning mate-pairs (clone coverage) as well as the average distance between the spanning mate-paired tags (average insert size).

Applying the look-up table and assessing statistical significance

The clone coverage at each genomic position (calculated above) is cross-referenced with the look-up table to determine the statistical significance of the observed difference in average insert size (compared to the expected value). Individual significance values for each library are combined by using a weighted sum. Combining multiple libraries generates a single weighted deviation at each genomic location.

Selecting candidate indels

Candidate indels are identified as regions in which each base pair is deviated by at least 3 standard deviations. Any regions that are within 2000 bp of each other and deviated in the same direction are combined into a single candidate indel. The clones spanning the coordinate with the most significant deviation are used to assess the candidate indel.

Hierarchical clustering to determine zygosity

The clones spanning the candidate indel are clustered according to the size of the clone when aligned to the reference sequence. Clustering continues until the difference in the sizes of all clones in a group is less than the range specified by the user.

Clusters containing less than two clones are eliminated. A single remaining cluster indicates a homozygous population, while two clusters indicate a heterozygous population.

Finally, the remaining 1 or 2 clusters are assessed for significance to determine if the candidate is called as an indel.

Post-processing

Data from individual chromosomes (processed separately) is combined.

3.2 Usage Parameters – Required

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

Parameter	Description	Type / Default / Example
-m, --mates-info	Full path to non-redundant mates file and information about the library paring range.	String and integer concatenated using a colon ':'. The library paring range is a library-specific parameter used to cluster individual clones. It can be easily estimated by subtracting the minimum insert size from the maximum insert

		<p>size of any given library. You may want to trim the range by eliminating highly divergent clones (outliers).</p> <p>Usage - Single library:</p> <pre>-m /path/to/F3_R3.mates.non-redundant:1000</pre> <p>Usage - Multiple libraries:</p> <pre>-m /path/to/F3_R3.mates.non-redundant:1000</pre> <pre>-m /path/to/F3_R3.mates.non-redundant:2000</pre>
<code>-r, --reference-cmap</code>	Full path to reference (*cmap) file.	<p>String.</p> <p>Usage:</p> <pre>-r /share/apps/corona/etc/cmap/human.cmap</pre>

3.3 Usage Parameters – Optional

Parameter	Description	Type / Default / Example
<code>-o, --output-dir</code>	Full path to the output directory where large indel analysis output will be stored. This directory should not preexist - it will be created during execution of the Large InDel tool.	String, Default = results/ Usage: <code>-o /path/to/results</code>
<code>-w, --warnings</code>	Show pre- and post-run warning messages. This parameter should be disabled if the Large InDel Tool is going to be integrated into a larger analysis pipeline.	Flag with no arguments. Default = Disabled Usage: <code>-w</code>
<code>-c, --max-clone-cov</code>	Maximum clone coverage for which the standard error (SE) will be calculated and stored in the look-up table.	Unsigned integer, Default = 1000 Usage: <code>-c 1000</code>
<code>-s, --min-stdev</code>	Minimum standard deviation from the expected (global) mean used to assess statistically significant insert size deviations and make indel calls. The value for this parameter depends on the size of the genome being analyzed. We recommend 6 for a human-sized genome.	Float, Default= 6 Usage: <code>-s 6</code>
<code>-n, --min-num-clust</code>	Minimum number of clones required to call an indel.	Unsigned integer, Default = 2 Usage: <code>-n 2</code>

3.4 Usage Example

```
large-indel-tool.pl -m ~/nova/large-indels/perl/t/test_F3_R3.mates.non-redundant:1000 -r ~/nova/large-indels/perl/t/chr1-4.cmap -q PBS -o ~/nova/large-indels/perl/t/test/results
```

3.5 List of programs/Scripts Included

- add_variant_number.pl
- apply_clone_lookup_table.pl
- assess_clusters_multiple.pl
- calculate_insert_average.pl
- calculate_physical_coverage.pl
- combine_assessments.pl
- combine_deviations.pl
- combine_indels.pl
- large_indel_cluster.pl
- parse_and_add_chromosomes.pl
- preproc.pl
- select_multiple_deviations.pl
- select_overlaps.pl
- table2gff.pl

3.6 Other Scripts Called by this Program

- submit_scripts_to_[PBS | LSF | SGE].pl
- Compute::Jobs.pm
- Compute::JobList.pm

3.7 Path Constraints

Input parameters should be the full path to the relevant files.

3.8 System Input Files

One non-redundant mates file per library and a single reference file and the corresponding single-entry fasta files.

3.9 Input File Versions Supported

File Type	Description	Fields
*.cmap	Tab-delimited file containing information related to the reference files. Each line represents a single reference with three required fields. Other fields can be added for application-specific functionality.	Each line represents a single reference with the following three required fields: <ol style="list-style-type: none"> 1. Reference index: From 1 to the number of chromosomes being analyzed. 2. Reference id: Usually a genome-specific chromosome identifier such as 22 or X. 3. The full path to the relevant reference file, represented in single-entry fasta format.

Example *.cmap files for several genomes can be found at `etc/cmap`.

3.10 Output Files

File Name (pattern) and location (path)	Description	Fields
<code><mates file name>.ref<id>AAA.AAB.AAC.distStartEnd</code> Location: <code>processed/</code>	Tab-delimited non-redundant mates file containing three extra columns (see Fields).	(1-11) Non-redundant mates file fields (12) Clone start position (13) Clone end position (14) Clone insert size
<code><mates file name>.AAA.lookup</code> Location: <code>processed/</code>	Tab-delimited look-up table calculated from AAA clones that maps each level of clone coverage to an associated standard error.	(1) Clone coverage (2) Global insert size (constant) (3) Global standard error at the given level of clone coverage.
<code><mates file name>.ref<id>.AAA.AAB.AAC.valid.physicalcoverage.insert_average</code> Location: <code>processed/stats/</code>	Tab-delimited file containing basic clone statistics at each genomic position.	(1) Clone coverage (2) Average insert size Each line is equivalent to a single genomic position (bp).
<code><mates.file.name>.ref<id>.AAA.AAB.AAC.valid.deviation</code> Location: <code>processed/stats/</code>	Tab-delimited output file generated by applying the look-up table to each genomic position.	(1) Genomic position (2) Clone coverage (3) Average insert size (4) Deviation from expected insert size (5) Number of standard errors from expected insert size. If the clone coverage at a particular genomic location is zero or exceeds that present in the look-up table, values of '-' are substituted for fields (4) and (5).
<code><mates file name>.ref<id>.combined.[beads clones clusters].<significance value></code> Location: <code>results/</code>	Files containing information about the clustering of clones spanning candidate indels.	Beads file: Contains clustered bead identifiers separated by colons ':'. Clusters file: Tab delimited file containing information about the cluster center (genomic location of maximum clone deviation), number

		of clones in the cluster, and associated bead identifiers in the cluster.
<pre>large-indels.[release gff]</pre> <p>Location: <code>results/</code></p>	<p>Tab-delimited text (release) or gff format final output file containing a list of indels ordered first by type (insertions then deletions), then by chromosome, and genomic position.</p>	<p>(1) Type of structural variation, either insertion or deletion, based on whether the average distance between mate-paired tags is less than or greater than the expected insert size, respectively.</p> <p>(2) Estimated upstream indel breakpoint.</p> <p>(3) Estimated downstream indel breakpoint.</p> <p>(4) Indel size, measured in bp.</p> <p>(5) Level of statistical significance.</p> <p>(6) Number of clones supporting the reference allele. Note: This value is only relevant for heterozygous loci—homozygous loci have a value of '-'. (7) Number of clones supporting the indel call.</p>

Several intermediate files are generated during large-indel detection and analysis. These files contain information needed to calculate position-specific statistics, determine statistically significant insert size deviations, and cluster similarly sized clones. The files are not needed to interpret the final output contained in the .gff file, but may be useful for understanding algorithmic aspects of the large-indel pipeline and considering candidate indels in technical detail.

3.11 Sample Input Files

```
nova/large-indels/perl/t/test_F3_R3.mates.non-redundant, nova/large-indels/perl/t/chr1-4.cmap
```

3.12 Sample Output Files

```
nova/large-indels/perl/t/bench/results/large-indels.gff
```

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.