SysCall Manual

1 Index

2 Running SysCall

Command:

```
perl SysCall.pl  list_candidates_heterozygous  reads_file.sam  outfile_base  code_dir
```

Example run:

After downloading SysCall, from within the directory './Example/' run:

```
perl ../SysCall.pl list_locations sample_reads.sam out_syscall ../
```

3 Input

Input: (1) list_candidates_heterozygous (2) reads_file.sam (3) outfile_base (4) code_dir

1. list_candidates_heterozygous

   This file is space delimited and holds in each row a candidate heterozygous position to classify. It also holds the nucleotides present around the location specified, 2 bases in each direction.

   Some example lines from hg18 build are:

   ```
   chr1 1840748 G G T A A
   chr1 17113024 G G A G C
   chr1 2467381 G G T C A
   chr1 17535338 T A C T T
   ```

Note: See appendix I for an explanation on how to easily generate the needed list file from a list of location coordinates.

Orientation note: If you do not use the instructions in Appendix I to generate your file please make sure to use the correct sequences in your list file. The nucleotides to be specified should always be of the forward strand of the published genome, such as, for example, specified on the UCSC Genome Browser when browsing those coordinates.

2. reads_file.sam

   This file contains the sequenced reads of your experiment in sam format. SysCall makes use of the binary score, location, sequence and quality scores at the reads.

   Some example lines are:

```
HWI-B5-690_0001:5:12:11067:13689#0 147 chr1 930038 255 76M = 929924 -190
ACCAGACCTCACTGTGTTGAAGTCATCGGCACCCCTTTCCTGCAGGAGGGGACACCTGCTCCCTGTCACTCTCCCG
AABDBABDBDADDDDBBDDCDCDCDBCCCC>C?BBBBBCC@CC@@CCCCCCDCDCCCCCCCCCCCCCCDCCCCCCCC XA:i:0 MD:Z:76 NM:i:0
HWI-B5-690_0001:5:14:16513:20412#0 83 chr1 930038 255 76M = 929924 -190
ACCAGCCCTCACTGTGTTGAAGTCATCGGCACCCCTTTCCTGCAGGAGGGGACACCTGCTCCCTGTCACTCTCCCG
####################################C-BB9CC9>C?B@>?@CBCCCBC?CCCCCCCCCCCCCCCC XA:i:1 MD:Z:5A70 NM:i:1
HWI-B5-690_0001:5:15:6549:10832#0 83 chr1 930038 255 76M = 929924 -190
ACCAGACCTCACTGTCTAGAAGTCATCGGCACCCCTTTCCTGCAGGAGGGGACACCTGCTCCCTGTCACTCTCCCG
A??B9<=>5>8.'*'++(8@=@2D@?>>;;543'&:ABB9BCC@@CCCCCCCCDCCCCCCCCCCCCCCCCCCCCC XA:i:2 MD:Z:15G1T58 NM:i:2
```

3. outfile_base

   Given the outfile_base, SysCall's outputs will be to the files out_syscall.heterozygous out_syscall.sys_errors. This will also be the basis for the names of the temporary files generated.

4. code_dir

   The directory to which SysCall was downloaded - the path you need to use to call SysCall.pl .

4 Outfiles

output: (1) list_heterozygous (2) list_sys_errors

1. out_syscall.heterozygous

   Will hold the lines from input file list_candidates_heterozygous that have been classified as heterozygous sites along with their posterior probabilities of being heterozygous sites.

2. out_syscall.sys_errors

   Will hold the locations from input file list_candidates_heterozygous that have been classified as systematic errors along with their posterior probabilities of being heterozygous sites.

Appendix I - Generating the candidate list file using the UCSC Table Browser.

In this section we give instructions on how to produce a candidate heterozygous file with the needed genomic coordinates.

1. Parse your list of candidate heterozygous sites into the following (tab or space delimited) bed format: <chr> <coordinate> <coordinate>.

   Example:

   ```
   chr20        61795        61795
   chr20        65493        65493
   chr20        65900        65900
   chr20        68179        68179
   chr20        68749        68749
   ```

   **notice that the same coordinate is specified in both columns.

2. Use the UCSC Table Browser to get the sequences around your locations:

   a. Go to the UCSC Table Browser at: `http://genome.ucsc.edu/cgi-bin/hgTables?org=Human&db=hg19&hgsid=193957763&hgta_doMainPage=1`

   b. specify the assembly you are using in the "assembly" field. **Make sure you have the correct assembly.**

   c. click on "manage custom tracks", upload your file in bed format (from step 1) and clique on "go to table browser".

   d. in the field "output format:" change to "sequence".

   e. at "output file:" specify a name for your outfile.

   f. press "get output"

   g. on the next page, titled "User Track Genomic Sequence" fill in the sentence with the numbers 3 and 2 as follows:
   Add 3 extra bases upstream (5') and 2 extra downstream (3')

   h. press "get sequence" to download your file.

   In the file downloaded the lines should look like this:

   ```
   >hg19_ct_UserTrack_3545_(null) range=chr20:68747-68751 5'pad=3 3'pad=2 strand=+ repeatMasking=none
   GTTGT
   >hg19_ct_UserTrack_3545_(null) range=chr20:69406-69410 5'pad=3 3'pad=2 strand=+ repeatMasking=none
   TCCGC
   ```

   Important Note : Make sure that the genome build is correct and that the middle location among the five bases is indeed the location you have specified as a candidate heterozygous site.

3. Now, parse the file output from the previous step into the format:
   <chr> <coordinate> <nuc-2> <nuc-1> <nuc> <nuc+1> <nuc+2>.

   Example:

```
chr20 61795 A G G A A
chr20 65493 T T A T A
chr20 65900 A C G T T
chr20 66052 T A G G A
chr20 68179 G T G T C
```

4. Congratulations - you are now ready to run the classifier!