CNV detection from the whole-exome sequencing data

Input/Output data files:

- regions_file (single file included with package)
- sample_ids (vector - one file for each sample)
- bam files (vector - one file for each sample)
- gender (required; 'U' if unknown)
- output folder (analysis folder for the batch)
- Read depth file names (vector - one file for each sample)
- Log2 ratio file names (vector - one file for each sample)
- MAD-weighted scores' file names (vector - one file for each sample)
- Sample Info File (single output file name to sample info)
- Break points file name
- Centromere regions file (included with package)
- Features file name

Short version:

CoNVex involves few simple steps:

(1) SamplePrepInfo() function prepares input/output files for analysis.
(2) Calculate read depth of probe regions (Agilent V3 probe regions file is bundled with R package and exists in ext_data folder) using java ReadDepth program that requires convex.jar, args4j-x.y.jar and sam-x.y.jar (also bundled with R package) in $CLASSPATH
(3) SampleLogRatioCallCommands() function creates a unix command to calculate log2 ratio
(4) BreakpointsCallCommands() function creates a unix command to calculate break points for number of reads (estimating regional MAD)
(5) GAMCorrectionCommands() creates multiple unix commands (one per sample) to correct log2 ratio and calculate MAD-weighted scores
(6) SWCNVCallCommands() creates multiple unix commands to call CNVs using Smith-Waterman algorithm.

For each R function and options, documentation is available in CoNVex.pdf file. All the above functions are wrapper scripts to create unix batch execution commands. You can also easily do it with 'awk' for example should you want it that way. Each step is dependent on the output of the previous step.

Long version:

Preparation of bait regions (hg19 Agilent V3 regions are included)

Genomic (exome + more) regions targetted by Agilent Sureselect 50mb library includes:

* CCDS exon regions
* GENCODE regions

* Non-coding RNAs
* etc.

It is also possible to design and target custom regions that are not covered by the Agilent library above (e.g., DDD exomes). These bait regions cover not only the actual target but also the regions nearby. For this algorithm, the nearby regions up to 100bp on either side of the target regions are included for the CNV detection as the coverage is high enough for the analysis.

Procedure:

Add 100bp to either sides of the bait region
Merge overlapping or gapless regions

This procedure is required for each new library. Whenever there is a new version of the Agilent library or a custom library is used in addition to the Agilent library, this procedure is repeated.

Sequence / Capture features

In order to estimate the systematic error introduced by factors both at the sequencing and capture stages, the algorithm uses the following:

* GC content
* deltaG (free energy of hybridization – between the RNA baits and the genomic DNA)
* Tm (Melting temperature – temperature at which half of the hybridization is complete)

deltaG and Tm are thermodynamic features that influence the hybridization of any two molecules – between the biotinylated RNA bait/probe and sheared genomic DNA in this case.

Read depth and #Reads

Two measures of coverage are calculated in this step:

- Mean read depth of each bait region above – this is the mean depth of all the bases in the region
- Number of reads falling in each regions

A java program does calculation of these two measures, which also takes care of pre-marked PCR/optical duplicates (most of the projects in Sanger mark them). However, if the PCR duplicates have been marked already, this step must be done explicitly.

Requirements:

- Picard Java API – to handle the bam files
- Args4j – to handle descriptive command line inputs

- Input files: bam files (and bai files – the bam index files either in the same folder (no action necessary), or in a different folder (must be explicitly specified in the command line options)
- Output: chr, start, end, #Reads, Mean depth (@MapQ >= 0)

Configuration Instructions:

* Download Sam-JDK – Picard Java API to manipulate bam files:
http://sourceforge.net/projects/picard/files/sam-jdk/

* Download args4j (an appropriate jar file):
http://args4j.kohsuke.org/

* Set your CLASSPATH environment variable (examples for *nix like environments below) to downloaded jar file(s). Use the latest version. Many older versions might work fine too!

```
export CLASSPATH=$CLASSPATH:/path/to/sam-x.y.jar:
:/path/to/args4j-x.y.z.jar:
```

* The Java program syntax (enter as single line):

```
java ReadDepth
-bam_file /path/to/Sample12345.bam
-regions_file /path/to/SureSelect_All_Exon_50mb.txt
-rd_file  /path/to/output/ProbeRD_Sample12345.dat
```

You can run it from the directory in which ExomeCoverage.class file exists – make sure current folder is in $CLASSPATH too. As long as ExomeCoverage.class is in $CLASSPATH, you can run the java program above from within any folder.

- regions_file having chr, start and end in the format below:

```
1    14539    15164
1    15571    16090
1    16491    18548
1    18949    19270
X    26491    28548
Y    28949    29270
```

* Output syntax:
chr, start, end, no. of sequencing reads overlapping that region, mean depth of all bases in the region @ MapQ >= 0

```
1    14539    15164    663  72.95
1    15571    16090    70   9.89
1    16491    18548    2784 100.8
1    18949    19270    320  70.98
```

Preparation of Samples' bam files

The preferred method of generating the read depth is to have one file for each sample in the project. This enables us to run the java program for each sample in an independent cluster node to parallelise the process.

The ReadDepthCommands() function in the CoNVex package creates a list of java commands – one for each sample – through the appropriate input and output files as described below:

Input parameters for ReadDepthCommands():

- regions_file containing the Chr regions (bait regions) – chr, start, end
- Sample IDs (vector)
- Bam files (vector)
- Bam index files (vector) – optional
- Output folder for read depth files
- Max memory for java program – default set to 1GB / required if the bam files quite big and use large memory
- More info:

```
?ReadDepthCommands()
```

For Agilent SureSelect All Exon 50mb library, the regions_file is already included with the R package (in the ext_data folder).