An Introduction to CnaGen

David Mosén-Ansorena

October 29, 2013

1 Introduction

1.1 Copy numbers and signals

A copy number alteration (CNA) is an acquired structural variation in which the number of copies of a certain region differs from the expected two copies in a diploid genome. For instance, one of the alleles can be duplicated in a region, yielding a copy number 3 alteration. Although a duplication of both alleles and the triplication of one of them result in a copy number 4 alteration, these are different cases: they produce different allelic imbalance. The cases where one of the alleles is missing are called loss of heterozygosity (LOH). LOH is acquired, or somatic, if the original diploid cells conserve both alleles. If the normal cells are already homozygous for the region, this is considered to be a germline LOH region.

An Illumina genotyping array provides two signals, which provide complementary information to characterize both copy numbers and allelic imbalances. These are the log R ratio (LRR), which reflects the total intensity signals for both alleles, and the B allele frequency (BAF), which is the relative proportion of one of the alleles with respect to the total intensity signal.

1.2 Issues

Tumour biopsies can be contaminated with normal cells whose genotypes are mainly diploid. This causes the LRR and BAF signals shrink and converge towards those of a diploid state proportionally to the degree of contamination. Also, tumours can be composed of subclones, this is, subpopulations of cells that harbour specific alterations along with the shared ones, which makes LRR and BAF signals even more complex.

Although each copy number has an expected LRR value and a specific BAF band pattern, these can be distorted by experimental probe-specific noise and by autocorrelated bias. The former affects both signals, whereas the latter affects only the LRR. Specifically, the autocorrelated bias is a consequence of genomic waves, caused by factors such as differences in GC content in the probes.

Figure 1 shows the BAF and LRR signals for some normal and altered regions. Notice the different copy numbers, allelic imbalances, contamination levels and how noise becomes a greater problem at greater contamination levels.

2 The package

CnaGen is an R package for the generation of synthetic SNP-array tumour samples with extensive parameterization. Normal cell contamination, intra-tumour heterogeneity, genomic waves, baseline shift and other known factors are parameterizable. In this document, we will focus on simple, illustrative examples, which will help you better understand how to define the parameters and thus to eventually generate data of your preference.

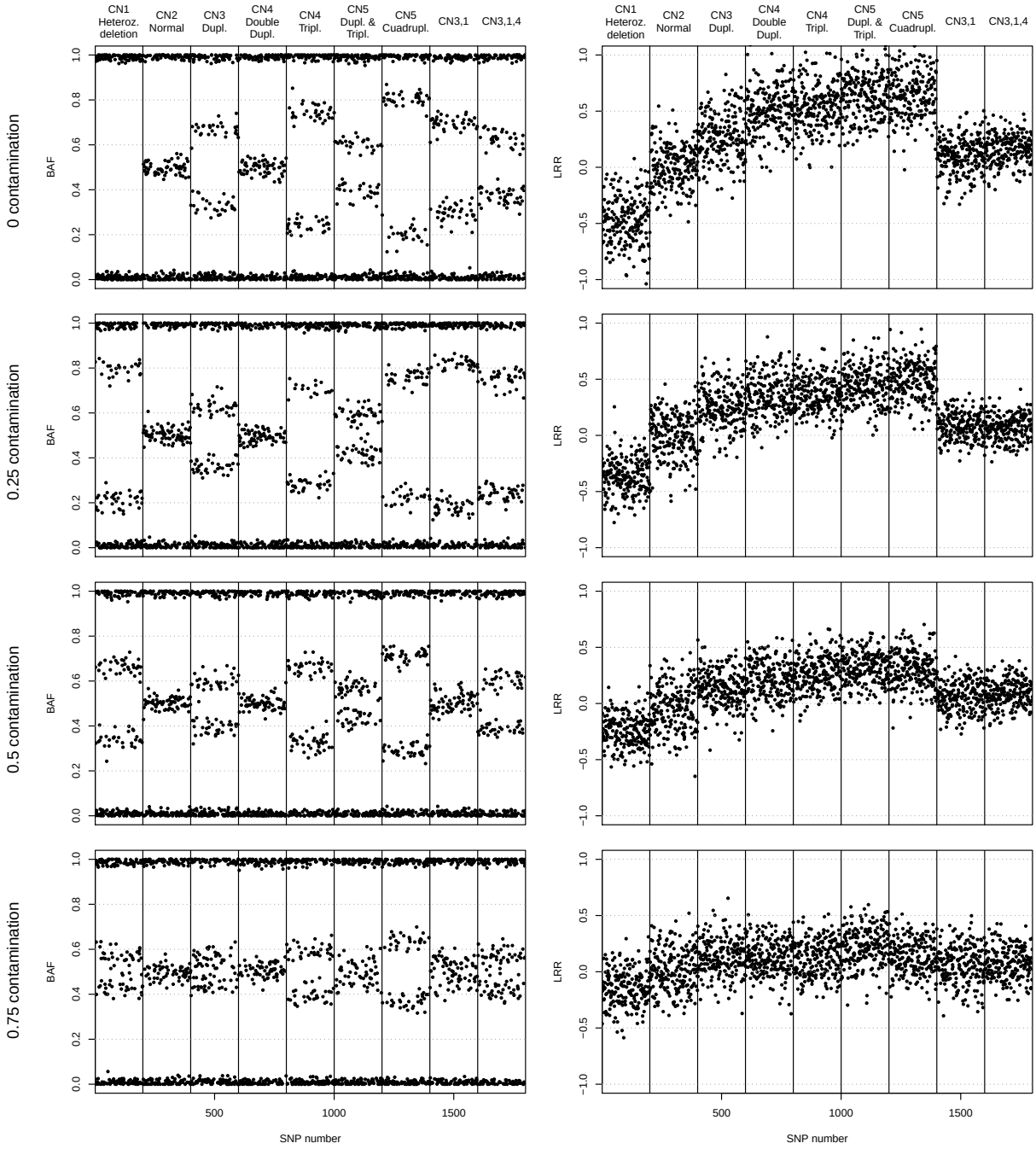

Figure 1: BAF and LRR signals of some example synthetic regions generated with CnaGen at different normal cell contamination levels: a heterozygous deletion (first column), a normal diploid region (second column), the various heterozygous CNA events up to copy number 5 (third to seventh columns) and two concrete cases of 2 and 3-subclone CNAs (last two columns).

The package is available at http://web.bioinformatics.cicbiogune.es/cnagen and to install it you should start R and enter:

```
install.packages("CnaGen_1.0.tar.gz", repos=NULL, type="source")
library("CnaGen")
```

2.1 The parameters

The parameters of CnaGen allow to taylor synthetic data according to the factors presented in the previous section.

The parameters 'lengths', 'times', 'weights2', 'weights3' and 'copynumbers' control the region-specific characteristics.

'noiseBAF', 'extremeBAF', 'reduceExtrBAF', 'noiseLRR', 'constLRRNoise' and 'waveSD' control the probe-specific noise and autocorrelated bias.

'baseMean', 'baseSD' and 'lrrLevels' adjust the levels for the copy numbers in the LRR signal. The baseline shift, controlled by the two former is correlated with the tumour purity of the samples, defined through the parameter 'purities'.

Certain SNP variations are more common than others. The frequency of the less common allele of a SNP is called PFB. In order to generate different PFB values for the synthetic SNPs, there are different options, provided by the parameter 'pfbMode'.

'annotationDir', 'lrrFirstDir' and 'bafFirstDir' allow you to choose where to output the synthetic samples and extra information on the generated data, such as the GC model and PFB files, and R data with region information. Set how many samples to output with the specified characteristics with the parameter 'replicates'.

Next, you will find a brief description of each parameter. The same list can be found in the help page of the cnagen method of the package, accessed typing in R ?cnagen.

- lengths: The set of lengths for the 6 region types (normal, simple CNA, somatic LOH, germline LOH, 2-subclone CNA, 3-subclone CNA). Use a list of vectors.
- times: For each region type, number of times that each unique region is included.
- weights2: Proportions that the main subclone will account for in 2-subclone CNAs.
- weights3: Proportions that the main subclone will account for in 3-subclone CNAs.
- copynumbers: Copy numbers for simple (normal and both LOH) regions, and vectors of copy numbers for 2-subclone and 3-subclone regions.
- replicates: How many times to generate samples at each of the specified purities.
- purities: Tumour purities at which to generate samples. Purity is the opposite of normal cell contamination.
- noiseBAF: Noise factor for the BAF signal. If left to 1, standard deviation will be 0.03.
- extremeBAF: Proportion of homozygous SNPs forced to take either 0 or 1 values in the BAF signal.
- reduceExtrBAF: Whether to reduce noise in the homozygous bands of the BAF signal.
- pfbMode: Select a PFB sampling mode from between 'array','threepeaks','uniform', 'maxinfo' and a value between zero and one.
- noiseLRR: Noise factor for the LRR signal. If left to 1 and constant standard deviation is set to TRUE, this will be 0.20.

- constLRRNoise: Whether to use the same standard deviation for all copy numbers in the LRR signal. If not, it decreases with copy number increase.
- lrrLevels: Expected LRR levels for the different copy numbers, starting with copy number 0.
- baseMean: How much to shift the LRR levels. Can be positive or negative.
- baseSD: Standard deviation for the variability in the LRR shift.
- waveSD: Standard deviation of the autocorrelated bias due to genomic waves, as described in Diskin et al., 2008. Default is 0.02.
- annotationDir: Output directory for the GC model and PFB files, and also additional R data that includes diploid genotype and region information.
- lrrFirstDir: Output directory for the samples in which the LRR column goes before the BAF column. Use them with OncoSNP.
- bafFirstDir: Output directory for the samples in which the BAF column goes before the LRR column.

3 Examples

3.1 Example 1

For starters, we will generate a short sequence with two complex alterations of 2 subclones (copy numbers 3 and 4) each.

```
cnagen(
  times = c(0,0,0,0, 1, 0),
  weights2 = 0.8,
  copynumbers = list(NULL,list(c(3,4)),NULL),
  purities = 0.5
)
```

The parameter 'times' commands how many times each region with an unique combination of factors appears. IT is a vector of length 6 because it can be separately defined for each of the 6 region types, in this order: normal, simple CNA, somatic LOH, germline LOH, 2-subclone CNA, 3-subclone CNA. In this example, only the 5th element of 'times' is set to 1, meaning that only regions with 2 subclones (5th region type) are included, and once per combination.

The parameter 'weights2' tells CnaGen the proportion(s) that the main subclone will account for, in this case 80%, and the parameter 'copynumbers' includes information on which subclones will comprise the alteration. Specifically, 'copynumbers' should be a list with 3 sublists. The first sublist will contain the copy numbers that simple regions (both CNAs and LOH regions) will take. The second and third list will contain the sets of copy numbers that will comprise the 2-subclone and 3-subclone CNAs. While the elements in the first sublist should be vectors of length 1, the elements of the second and third list should be vectors of lengths 2 an 3, respectively.

Two 2-subclone alterations will be generated in this example, one in which the copy number 3 subclone accounts for 80% of the intra-tumour mix and another one in which the copy number 4 subclone will. Notice that 'weights2' can be a vector, so more combinations of alterations could have been generated. If you want to include 3-subclone regions, you should tweak the last element of 'times' and the last sublist of 'copynumbers', but also include the 'weights3', which works similarly to 'weights2'.

Because the purity (1 - contamination) of this synthetic data is set to 50% with the 'purities' parameter, the actual proportion of the main subclone will be 40%. Again, 'purities' can be a vector, so several synthetic samples with varying normal cell contamination can be generated at once.

Here is a possible output sequence:

```
   start end   proportions copy.numbers homozygosity length main.cn main.cn.prop
 2     1 100 0.5, 0.1, 0.4      2, 3, 4           no    100       4          0.4
 1   101 200 0.5, 0.4, 0.1      2, 3, 4           no    100       3          0.4
```

If you want to review this kind of table for your own data, you can find the corresponding R object in the directory specified by the 'annotationDir' parameter. Look for a folder named 'data' in your working directory if you did not use such parameter.

3.2 Example 2

The elements in parameter 'lengths' can also be vectors, so that regions of each type are generated with varying lengths. In this case, we generate 6 simple CNA regions of 3 different lengths and 2 different copy numbers. Notice how the first sublist of 'copynumbers' is filled with 2 separate vectors of 1 element each:

```
cnagen(
  lengths = list(NULL, c(10,20,50), NULL,NULL,NULL,NULL),
  times = c(0, 1, 0,0,0,0),
  copynumbers = list(list(c(3),c(4)),NULL,NULL),
  purities = 1
)
```

Here is a possible outcome:

```
 Possible output regions:
   start  end proportions copy.numbers homozygosity length main.cn main.cn.prop
 2     1   20        0, 1         2, 3           no     20       3            1
 5    21   40        0, 1         2, 4           no     20       4            1
 6    41   90        0, 1         2, 4           no     50       4            1
 3    91  140        0, 1         2, 3           no     50       3            1
 4   141  150        0, 1         2, 4           no     10       4            1
 1   151  160        0, 1         2, 3           no     10       3            1
```

3.3 Example 3

In this last example with the most intrincate parameters, we will generate 2 normal diploid region of 50 and 100 SNPs and 2 somatic LOH regions (copy numbers 1 and 3) of 10 SNPs repeated twice each.

```
cnagen(
  lengths = list(c(50,100), NULL, c(10), NULL,NULL,NULL),
  times = c(1, 0, 2, 0,0,0),
  copynumbers = list(list(c(1),c(3)),NULL,NULL),
  purities = 0.2
)
```

Possible output regions:

```
   start  end proportions copy.numbers homozygosity length main.cn main.cn.prop
 1     1   50           1            2           no     50       2          1.0
 3    51   60    0.8, 0.2         2, 3      somatic     10       3          0.2
 2    61   70    0.8, 0.2         2, 1      somatic     10       1          0.2
 5    71   80    0.8, 0.2         2, 3      somatic     10       3          0.2
 4    81   90    0.8, 0.2         2, 1      somatic     10       1          0.2
 6    91  190           1            2           no    100       2          1.0
```

3.4 Example 4

With this example, you will learn how to use the basic parameters associated to technical and biological biases and noises. The following call to cnagen produces BAF and LRR signals without noise, as shown in Figure 2.

```
cnagen(
lengths = list(c(200), c(200), NULL,NULL, c(200), NULL),
times = c(1, 1, 0,0, 1, 0), weights2 = 0.8, purities = 1,
copynumbers = list(list(c(1),c(3),c(4)), list(c(3,4)), NULL),

noiseBAF = 0,
noiseLRR = 0,
waveSD = 0
)
```

If we change the 'waveSD' to 0.03, we are adding genomic waves to the LRR signal, which will look like Figure 3. If you are not sure which values to assign these noise parameters, the genomic waves have been found to have a standard deviation between 0.01 and 0.04. In turn, if you set the probe-specific noises of BAF and LRR to 0.03 and 0.2, respectively, CnaGen will produce data with noise levels similar to those found in real samples.

3.5 Example 5

In this final example, you will see how the 'extremeBAF' and 'reduceExtrBAF' parameters affect the homozygous bands of the BAF signal. The code below was used to generate the data whose BAF signal appears in the first column of Figure 4.

```
cnagen(
```

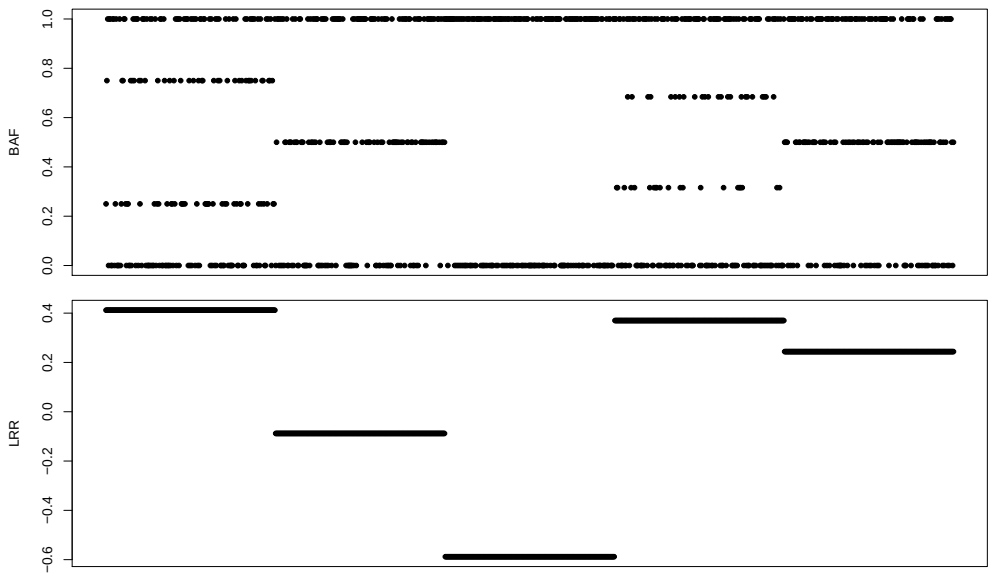

Figure 2: BAF and LRR signals of some synthetic regions without noise.

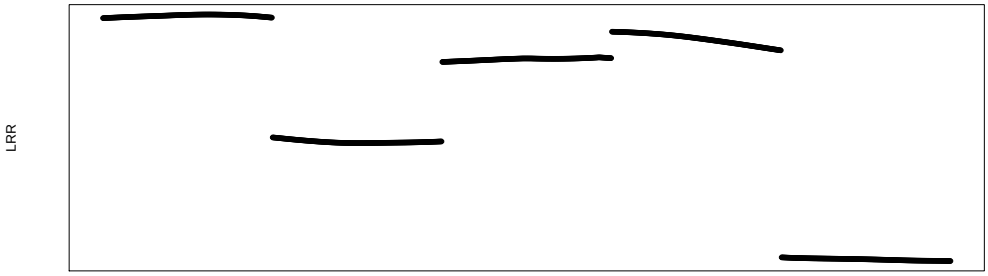

Figure 3: LRR signals of some synthetic regions without probe-specific noise but with the bias produced by genomic waves.

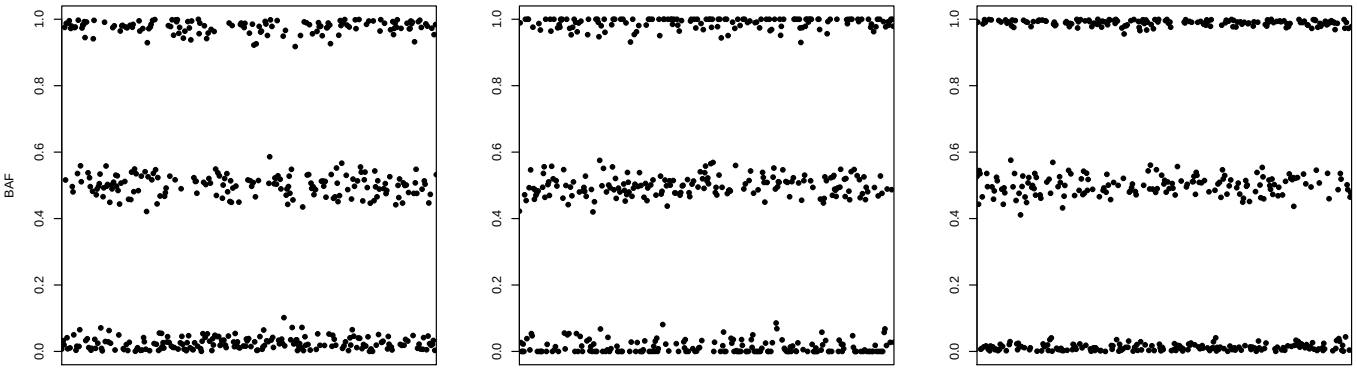

Figure 4: Variations in the BAF signal's homozygous bands.

```
lengths = list(c(500), NULL,NULL,NULL,NULL,NULL),
times = c(1, 0,0,0,0,0),
copynumbers = list(NULL,NULL,NULL),
purities = 1,
extremeBAF = 0,
reduceExtrBAF = FALSE
)
```

In the second column of Figure 4, the homozygous bands of the BAF signal were forced to take the values 0 or 1 in half of the cases, giving the 'extremeBAF' parameter a value of 0.5. In the third column, these bands have a reduced noise, specifically half of the heterozygous band, by setting the 'reduceExtrBAF' parameter to TRUE.