FG - Assignment 2 - 2018
Transforming relative copy number to absolute

16/11/2018

The goal of this assignment is to develop a method for transforming relative copy number profiles to abosolute copy number profiles. As with the development of any new bioinformatics method, you will first simulate data to generate a ground truth. The simulated data will be used to test and optimise your method. You will then apply the method to real data. You can assume that all simulated and real samples are human tumour samples and they are 'pure' (i.e. are not contaminated by normal cells). Therefore, you only need to scale (linearly) the relative copy number to get absolute copy number profiles.

1 Instructrions

- You are required to submit two files:
  - fga2_XXX.Rmd (where XXX is your CRSid). This should be a Rmarkdown file containing all source code which when compiled generates a pdf of your submission.
  - fga2_XXX.pdf (where XXX is your CRSid). This should be your submission pdf.
- The due date is: November 30, 2018.
- This assignment comprises 30% of your overall mark
- You will be assessed on your ability to complete all tasks, accurately display results, clearly document code, assess performance of method, interpret results, and justify your approach.

2 Write a copy number profile simulator

Write a simulator to generate human genome tumour copy number profiles. Use your simulator to simulate 5 copy number profiles with 3, 10, 25, 100, 200 copy number changes. Only simulate single copy gains or losses. Events can range from 1MB to 100MB in size. Plot the copy number profiles.

Hints and assumptions: You only need to simulate segmented copy number (not underlying read/SNP data). Events cannot be larger than a chromosome (use the hg19.chrom.sizes.txt annotation file to determine chromosome sizes). Copy number changes happen sequentially, therefore one change can occur on top of another. If a change does occur on top of another, it can only be as large as the segment in which it occurs (e.g. an event cannot span multiple segments). The initial genome should be diploid. Don't worry about allele specific copy number (only output total copy number). A suggested tabular format for a copy number profile is: chromosome, start, end, copy-number

3 Generate copy number profiles with noise

Similar to above, simulate 5 copy number profiles with 3, 10, 25, 100, 200 copy number changes. This time, introduce segmentation noise and allow copy number changes up to 2 copies.

Hint: rather than sampling from integer copy number changes, sample from a gaussian centered at an integer state and introduce noise by setting the standard deviation to 0.1.

4 Generate copy number profiles with a tetraploid background

Similar to above, simulate 5 copy number profiles with 3, 10, 25, 100, 200 copy number changes, however, this time start from a tetraploid genome rather than a diploid genome.

5 Develop a method for assessing the difference between a copy number profile and integer copy number states

Develop a measure which determines how close the copy-number profile is to a 'clonal' profile (i.e. copy number states that are integer values).

6 Generate a relative copy number profile

Simulate a genome with 10 single copy changes with no noise. Convert the profile to relative copy number by median normalising and plot.

7 Transform relative copy number profile to absolute

Chose a range of scaling factors to convert the relative copy number profile above into an absolute copy number profile. Calculate your measure for each scaling factor and plot. Does your measure indicate the correct solution? Do you observe any interesting patterns? Comment.

8 Transform relative copy number profile to absolute (with noise)

Repeat above this time introducing segmentation noise. Plot and comment.

9 Apply method to real data

Determine the optimal scaling factor for transforming the relative copy-number profiles found in "relative_segment_tables.rds". Plot the absolute copy number profiles. Justify your selection of scaling factor.

Hints: use loadRDS to read in the relative copy number profiles. The profiles are stored as a list of segment tables.