

Readme of R programs for the Bayesian Segmentation Approach

January 12, 2009

Introduction

In this document, we provide the usage of R programs for implementation of Bayesian segmentation approach (BSA) (Wu *et al.* 2009) in analysis of the HapMap data (<http://www.sanger.ac.uk/humgen/cnv/>) at the population level. In addition to proper setup of the priors of expected intensity ratios at different copy number levels and the stopping criterion c , successful application of these programs requires proper data preparation, referred to as the “vectorization” process of the original data matrix.

In the following sections, we first give a list of R programs, then give an example of using the program and explain the outputs.

List of programs

This set of R programs are grouped into three files

- BSA_Library.txt
- WGTP_DUP.txt
- WGTP_DEL.txt

The BSA_Library contains the programs that are needed by WGTP_DUP or WGTP_DEL. Both WGTP_DUP and WGTP_DEL are similar programs but with different prior values for identifying duplications or deletions. For details of the algorithm, please consult Wu *et al.* (2009). To analyze a dataset, the BSA_Library needs to be loaded into R first, then the WGTP_DUP or WGTP_DEL.

Data preparation

We give an example of “vectorization” of a data matrix. The following data, in terms of data matrix, are extracted from the \log_2 ratios of chromosome 17 of the CEU population. The first column is BAC clone names; the second chromosome name; the third and the fourth the starting and ending genomic locations of the corresponding BAC clone. NA12144, NA12145, NA12146, NA12239, NA06994, and NA07000 are individual ID numbers for six CEU individuals.

clone	chr	start	end	NA12144	NA12145	NA12146	NA12239	NA06994	NA07000
Chr17tp-5D2	17	343377	440727	-0.0705	-0.0380	-0.0575	0.0710	0.1470	-0.0090
Chr17tp-3F6	17	420484	585964	0.0340	-0.0045	-0.0615	-0.0070	0.0550	-0.0055
Chr17tp-7A1	17	568336	736705	-0.0095	-0.0135	0.0120	-0.0260	0.0200	-0.0805
Chr17tp-3C8	17	707754	880135	0.0325	0.0760	-0.0020	0.0575	0.0235	-0.0650
Chr17tp-2F1	17	800495	1008155	-0.0045	NA	0.0895	0.0125	0.0230	0.0025

We first remove the missing observation. After “Vectorization”, the above data should be in the following format

start	log2ratio
343377	-0.0705
343377	-0.0380
343377	-0.0575
343377	0.0710
343377	0.1470
343377	-0.0090
420484	0.0340
.	.
.	.
.	.
800495	0.0025

Note that we only use the starting positions of the clones. The complete genomic locations can be reconstructed using these starting positions from the original data matrix. Data in this format will be used in the programs of WGTP_DUP to identify copy number duplications. To identify the deletions, we multiply the \log_2 ratios by -1 and feed the modified data to the program WGTP_DEL. This “vectorization” process is relatively straight forward and can be done more efficiently in other programs such as PERL or SAS, therefore, the R script is omitted. For running the program, i.e. WGTP_DUP, the following command is issued in R:

```
WGP_DUP(x=start, y=log2ratio, stop.value=0.1)
```

The stop.value is the stopping criterion c .

Program output

We feed the “vectorized” data of chromosome 17 of the CEU population into WGTP_DUP, the outputs are as follows,

chr.names	Start	End	Num_of_obs	Average	Variance	Bayes_Factor	Eta	Eta0	a	c_ratio
17	41559185	41747597	235	0.2044	0.0321	6.7679	0.25	0	5.6537	0.1769
17	41798857	41823594	115	0.1609	0.0818	3.1190	0.25	0	16.3819	0.0610

The **chr.names** indicates the chromosome name. The **Start** and **End** locations define a segment. Note that both locations are actually the starting positions in the original data matrix. The **End** positions can be used

as the search key in the original data matrix to infer the correct ending positions of the identified segments. The **Num_of_obs** indicates the number of observations in the corresponding segments. **Average** and **Variance** are the sample mean and variance of the \log_2 ratios of the segment. **Bayes.Factor** indicates the Bayes factor values of the identified segments. **Eta** is the a_0 value and **c_ratio** is $\frac{1}{a_0}$, see the supplementary information of Wu *et al.* (2009) for more details of a_0 .

References

Wu, L.Y., Chipman H., Bull, S.S., Briollais, L., Wang K. (2009) A Bayesian segmentation approach to ascertain copy number variations at the population level. submitted.