Published in final edited form as:

IEEE Int Conf Bioinform Biomed Workshops. 2010 December ; 2010: 827–828. doi:10.1109/BIBMW.2010.5703928.

Correction of Copy Number Variation Data Using Principal Component Analysis

Jiayu Chen[1], **Jingyu Liu**[1,2], and **Vince D. Calhoun**[1,2]

[1]Dept. of Electrical Engineering, University of New Mexico, Albuquerque, NM

[2]The Mind Research Network, Albuquerque, NM

Abstract

Copy number variation (CNV) detection using SNP array data is challenging due to the low signal-to-noise ratio. In this study, we propose a principal component analysis (PCA) based correction to eliminate variance in CNV data induced by potential confounding factors. Simulations show a substantial improvement in CNV detection accuracy after correction. We also observe a significant improvement in data quality in real SNP array data after correction.

Keywords

copy number variation; Log R Ratio; principal component analysis

I. INTRODUCTION

Copy number variation (CNV) is one type of genetic variation caused by large segmental insertions or deletions of DNA sequence. One often used technology to assess genomic CNVs is through single nucleotide polymorphism (SNP) arrays. Typically, the log R ratio (LRR) and B allele frequency (BAF) are measured for each SNP locus. Our research explores the use of principal component analysis (PCA) to eliminate variance in LRR data induced by potential confounding factors, thus enhancing the validity of CNV detection.

II. METHODS

Through PCA, an LRR data matrix (genetic loci-by-samples) is decomposed into a linear combination of underlying principal components (PCs) or sources, as shown in (1). Each PC accounts for a certain amount of variance of the data. Pearson correlation or analysis of variance (ANOVA) is used to assess the associations of PCs with potential continuous or categorical confounding factors, respectively. The association is evaluated either in the genomic loci dimension such as with GC-percentage, or in the sample dimension such as with batch effect. A PC (e.g., the k^{th} component) with a significant association after Bonferroni correction is identified to represent a confounding factor and needs to be removed for correction, as illustrated in (2) and (3).

$$X = \sum_{i=1}^{r} u_i \sigma_i v_i^T \quad (1)$$

$$X_k = u_k \sigma_k v_k^T \quad (2)$$

$$X_c = X - X_k \quad (3)$$

Synthetic SNP array data were prepared to evaluate the effectiveness of the PCA-correction. To closely represent the real data characteristics, we inherited the chromosome 1 markers' names and positions in the Illumina Human-1M Duo SNP array (97964 markers), and simulated three types of noise effect: GC-percentage [1], batch effect (scanner and processing date) and random Gaussian noise. A total of 200 samples were generated.

The number of bad samples that failed quality control (standard deviation of LRR, LRR_SD, < 0.28) [2], was employed as a measure of data quality for both simulated and real SNP array datasets. Additionally, we applied PennCNV [3] to the simulated dataset and computed the false positive rate (FPR) and false negative rate (FNR) of CNV calls, as indices of detection accuracy by comparing the PennCNV results with the ground truth.

III. RESULTS

Ten independent simulated datasets were tested and results showed consistent performance. In PCA-correction, the first two components were identified to reflect GC-percentage and batch effect and then removed, as summarized in Table 1a. After correction, the data quality was able to improve dramatically, as shown in Table 1b. Particularly, the number of bad samples decreased from 76 to 40 in high noise group and 10 to 4 in low noise group. The accuracy of CNV detection, consequently, was improved. In Table 1c, we can see a significant improvement of FPR in CNV detection, as well as a slight improvement of FNR. Finally, a performance comparison was made between PCA-correction and regression-based correction [1], as shown in Table 1d. The results indicated comparable detection accuracies, with PCA-correction showing slight improvements.

To separately investigate the influences of different types of noise on CNV detection, we hereafter only use the correction for GC-percentage, and the corrected data refer to the data after GC-percentage correction.

ANOVA test indicated a significant group difference between high and low Gaussian noise, in terms of false negatives (FNs), as illustrated in Table 2. A further comparison was made among three independent datasets, different only in the level of GC-percentage effect, measured by the absolute correlation between GC-percentage and the simulated LRR data ($|r_{GC\text{-LRR}}|$). The results indicated that false positives (FPs) were greatly influenced by GC-

percentage effect, with ANOVA test showing significant group differences, as shown in Table 3. However, after the PCA-correction for GC-percentage, the FPRs all went down to a low level around 0.04 and no group difference was observed.

The quality of real SNP data was able to improve dramatically as well after correction. Table 4 lists the PCA results and the data quality evaluation. The median LRR_SD decreased from 0.24 to 0.18 with 31 more samples saved.

IV. DISCUSSIONS AND CONCLUSIONS

PCA-correction:

The main advantage of PCA-correction is that it provides a complete data decomposition, which allows simple and non-parametric data correction for any type of confounding factor. Both simulation and experimental results show a substantial reduction in data variance after correction. Simulations further show that PCA-correction can help CNV detection by significantly reducing FPR and slightly reducing FNR. Compared to the regression-based method, PCA-correction can be flexibly extended to correct other categorical confounding factors along sample space, such as batch effect, whose influence on the data may be difficult to isolate otherwise.

GC-percentage vs. Gaussian noise:

These two factors influence differently two types of error. The GC-percentage induces a wave pattern in the LRR data. Since the CNV detection is to locate regions with significant alternations in LRR, it is naturally sensitive to waviness, which explains the higher FPR when GC-percentage effect is stronger, as shown in Table 3. On the other hand, with increased variation induced by Gaussian noise, the difference in LRR between aberrant and normal regions becomes less significant, which leads to increased false negatives, as indicated in Table 2.

In summary, we propose a PCA-based correction for LRR data. Both simulation and experiment results show that PCA-correction significantly decreases the fluctuations in LRR data, and simulations further confirm that PCA-correction leads to a significant improvement in FPR and a slight improvement in FNR. Overall, PCA-correction is designed to work with existing CNV detecting algorithms to enhance the validity CNV calls.

ACKNOWLEDGMENT

This work was supported by grants from the National Institutes of Health (1R21DA027626 & R01EB005846).

REFERENCES

[1]. Diskin SJ et al. “Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms”, Nucleic Acids Res, vol. 36, e126, Sep. 2008 [PubMed: 18784189]

[2]. Need AC et al. “A genome-wide investigation of SNPs and CNVs in schizophrenia”, Plos. Genet, vol. 5, e1000373, Feb. 2009 [PubMed: 19197363]

[3]. Wang K et al. “PennCNV-An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data”, Genome Res, vol. 17, pp. 1665–1674, Oct. 2007 [PubMed: 17921354]

Table 1.

PCA and data quality evaluation

Table 1a. PCA-correction				
Designed feature	GC-percentage		Date and scanner	
Identified Component	1st		2nd	
P-value	<1E-23		<1E-23	
Table 1b. Evaluation of data quality				
Data Quality	High noise		Low noise	
	σ_{LRR}	N_{sub_ex}	σ_{LRR}	N_{sub_ex}
Uncorrected	0.30±0.03	76	0.25±0.03	10
Corrected (Comp. 1)	0.28±0.02	46	0.23±0.02	4
Corrected (Comp. 1, 2)	0.28±0.02	40	0.22±0.02	4
Table 1c. Detection Accuracy: PCA-correction				
Total generated markers with CNVs: 75867				
PennCNV results	Overall FPR		Overall FNR	
Uncorrected	0.6220		0.1374	
Corrected (comp. 1)	0.0389		0.0940	
Corrected (comp. 1, 2)	0.0351		0.0886	
Table 1d. Detection Accuracy: regression-based correction				
PennCNV results	Overall FPR		Overall FNR	
GC-percentage corrected	0.0389		0.0944	

Note: High noise and low noise refer to the groups with high-SD and low-SD Gaussian noise, each group containing 100 samples; N_{sub_ex} denotes the number of bad samples that failed quality control; FPR and FNR are calculated using the total number of markers with CNVs.

Table 2.

Evaluation of false negatives vs. Gaussian noise

Corrected	High noise	Low noise
$\sigma_{Gaussian}$	0.28±0.02	0.22±0.01
FNs	21±32	6±11
ANOVA	P-value = 4.92E-06	

Table 3.

Evaluation of false positives vs. GC-percentage

Uncorrected	**GC1**	**GC2**	**GC3**
$\lvert r_{LRR\text{-}GC} \rvert$	0.35±0.21	0.30±0.18	0.25±0.17
FPs	444±678	236±407	155±334
ANOVA	P-value = 2.34E-08		
Overall FPR	GC1	GC2	GC3
Uncorrected	1.1710	0.6220	0.4090
Corrected	0.0413	0.0389	0.0317

Note: GC1, GC2 and GC3 represent the three datasets with different levels of GC-percentage effect, each dataset containing 200 samples.

Table 4.

PCA and data quality evaluation

Factor	Component	P-value
GC-percentage	1^{st}	<1E-23
Gender	15^{th}	6.31E-11
Data Quality	Uncorrected	Corrected
σ_{LRR}	0.24 ± 0.08	0.18 ± 0.04
N_{sub_ex}	37	6