Reducing system noise in copy number data using principal components of self-self hybridizations

Yoon-ha Lee*, Michael Ronemus*, Jude Kendall†, B. Lakshmi†, Anthony Leotta‡, Dan Levy+, Diane Esposito*, Vladimir Grubor*, Kenny Ye†, Michael Wigler*#, and Boris Yamron*#

*Cold Spring Harbor Laboratory; 1 Bungtown Road, Cold Spring Harbor, NY 11724; and ‡Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461

AUTHOR SUMMARY

Genomic copy number variation (CNV) is a phenomenon wherein the number of copies of a certain gene varies among individuals, causing major differences in individual characteristics and resulting in widespread genetic disorders (1). CNVs can be detected by hybridizing genomic DNA to fragments of nucleic acid (i.e., probes) used to detect target nucleotide sequences in samples (2, 3). One common method, comparative genomic hybridization (CGH), is useful but often produces irrelevant background signals, or noise. Here, we use a method known as principle component correction (principal component correction) to enhance signal in the presence of noise, improving results and reducing costs.

In CGH, two genomes—an experimental sample and a reference—are hybridized to a microchip consisting of many probes. Analysis of the CGH ratio data can reveal trends in genomic regions shared by some hybridizations and not by others (Fig. P1A). These trends often interfere with CGH segmentation analysis, a common method for analyzing the genome, and lead to spurious signals. When a pair of hybridizations shares trends in one region of the genome, that pair typically has long-range correlations throughout the genome. Such correlations between the genomes of unrelated individuals conflict with expectations based on genetic laws. Trends observed in test data are often present even when DNA samples from the same genome are repeatedly hybridized (i.e., self-self hybridizations), leaving no doubt that these “trends” are noise rather than true genetic signal (Fig. P1A).

Not all trend patterns are alike, but instead are composed of relatively independent components. One major trend is associated with the content of two nucleotide bases (GC content) (4), but it is not the only one. We sought to correct for correlated system noise while minimizing overall adjustment. First, the major (low-dimensional) orthogonal basis (PC) for the system noise was found using principal component analysis. Second, we maximized the fit of sample ratio data to the PC using the least squares method, and considered the residual signal as the true genetic signal. To avoid mixing true genetic signal with system noise in the principal component analysis, we built the noise basis from analysis of self-self hybridizations, which contain no genetic signals, only noise. The results of such corrections are illustrated in Fig. P1B. This is what we call principal component correction.

Principal component correction improves CGH by reducing trends and long-range correlations in the data, improving signal-to-noise metrics, and reducing false segmentation. When principal component correction was applied to our analysis, all test hybridizations (100%) had decreased total noise, and 91.51% of our hybridizations had decreased autocorrelation, a direct measure of trends. The mean relative improvement [100 x (before-after)/before] of total noise was 11.2%, and the mean relative improvement of the autocorrelation was 33.1%. This method does not introduce detectable spurious signals, which would otherwise result from using actual test data to form PCs.

We also tested a variant of our standard procedure to deal with components of noise that were readily detected by principal component analysis, but not corrected well by principal component correction. Rather than treat all hybridization probes equally, we partitioned the probes into those most sensitive to particular components of system noise and then separately used principal component correction to correct these probes. We call this “piecewise principal component correction,” and it improves correction of the “hard-to-treat” noise components.

In addition to enabling subtraction of system noise, the PCs themselves provide critical insights into the sources of this noise. In our system, the loadings of the PCs correspond to known probe variables, such as discrete physical location of the probes on the microarray surface used for the


The authors declare no conflict of interest.

This Direct Submission article had a prearranged editor. Freely available online through the PNAS open access option.

Data deposition: Raw and processed data files corresponding to all hybridizations in this study have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE23682).

1Present Address: Ontario Institute for Cancer Research, Toronto, Ontario, Canada MSG 0A3.

2Present Address: Institute for Genome Science and Policy, Duke University, Durham, NC 27708.

To whom correspondence may be addressed. E-mail: wigler@cshl.edu.

See full research article on page E103 of www.pnas.org.

Cite this Author Summary as: PNAS 10.1073/pnas.1106233109.
hybridization, composition of DNA in terms of the nucleotide bases it contains, and proximity of the probes to genes. Joint analysis of test data and the PCs also reveals operational variables. Finally, analysis reveals some inadequacies of the CGH correction, and points to regions of the genome prone to errors—which may result from the structure of chromatin, the complex of proteins and DNA that packs the genome into the nucleus.

Our new method offers a number of advantages, giving it the potential to improve genetic hybridization analyses while reducing costs. It adds some minimal expense (less than 5% of the total), but provides results comparable to or better than those of common expedients, such as color reversal or technical replicates, nearly halving the total study cost overall. Work in progress indicates that our method, using repetitions of a single reference, can be used to suppress noise even in data derived from a single channel of a microarray. This can again halve the cost of assay. We expect that our technique should be generally applicable to genomic copy number data gathered by any method, including DNA sequencing, a technique now indispensable in biotechnology, diagnostics, and many other fields.