Topic Introduction

Copy Number Variation Detection via High-Density SNP Genotyping

Kai Wang¹ and Maja Bucan

Department of Genetics, University of Pennsylvania, Philadelphia, PA 19446, USA

INTRODUCTION

High-density single nucleotide polymorphism (SNP) genotyping arrays recently have been used for copy number variation (CNV) detection and analysis, because the arrays can serve a dual role for SNP- and CNV-based association studies. They also can provide considerably higher precision and resolution than traditional techniques. Here we describe PennCNV, a computational protocol designed for CNV detection from high-density SNP genotyping data. This protocol extracts allele-specific signal intensities from genotyping arrays, and then integrates information on SNP spacing and SNP allele frequencies to generate CNV calls by a hidden Markov model (HMM) algorithm. Analyses of CNVs from SNP genotyping arrays will provide a more comprehensive view of genome variation, and complement current genome-wide association studies in identifying disease susceptibility loci.

OVERVIEW

CNV refers to genomic segments of at least one kb in size, for which copy number differences have been observed in comparison to reference genome assemblies (Feuk et al. 2006; Freeman et al. 2006). Multiple large-scale studies have reported prevalent CNVs in humans, suggesting that they may account for a significant portion of phenotypic variation (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005; Conrad et al. 2006; McCarroll et al. 2006; Redon et al. 2006). The precise and comprehensive identification of CNVs would greatly benefit the functional analysis of human genome variation, and complement current genome-wide association studies that use SNPs.

In recent years, several techniques have been developed to detect CNVs in human and other mammalian genomes (Carter 2007). Traditionally, large chromosome rearrangements have been detected with G-banded karyotypic analyses or fluorescence in situ hybridization (FISH) using fluorescent probes that bind to part of the chromosomes. With the development of microarray technology, array-comparative genomic hybridization (CGH) platforms can be used to detect gains and losses of genomic segments (Pinkel et al. 1998), especially in tumor tissues. This technique involves labeling a reference genome and a testing genome with different fluorescence markers, hybridizing to the microarray with genomic clones, and then analyzing the intensity of the hybridization signal for each clone. Further variations of the array-CGH technique have been developed to detect CNVs in human populations (Iafrate et al. 2004; Sebat et al. 2004; Locke et al. 2006; Redon et al. 2006). Genome-wide oligonucleotide arrays use short oligonucleotides spotted to arrays by inkjet technology or by photolithography. They offer higher resolution than large-insert clone array platforms and can achieve complete genome coverage. Furthermore, paired-end mapping (Korbel et al. 2007) and clone-end resequencing (Eichler et al. 2007) have also been implemented for CNV detection, and can identify many more small-scale CNVs. Finally, whole-genome resequencing (Service 2006) could soon prove cost-effective for the most comprehensive identification of CNVs in the human genome.

¹Corresponding author (kai@mail.med.upenn.edu)
Please cite as: CSH Protocols; 2008; doi:10.1101/pdb.top46

In addition to the techniques mentioned above, high-density SNP genotyping arrays have become more popular for copy number detection and analysis, because the arrays can serve a dual role for both SNP-based and CNV-based association studies. In high-density SNP genotyping platforms, a signal intensity measure is summarized for each allele of a given SNP marker. Analysis of signal intensities across the genome can then be used to identify regions with multiple SNPs that support deletions or duplications (Komura et al. 2006; Peiffer et al. 2006). However, there are limitations to the use of SNP genotyping arrays for CNV detection: SNPs in these arrays are not uniformly distributed across the genome and are sparse in regions with segmental duplications or complex CNVs (Carter 2007). To overcome these limitations, the new generations of SNP genotyping arrays from Affymetrix and Illumina have now incorporated additional nonpolymorphic (NP) markers to provide more comprehensive coverage of the human genome. We describe here a computational protocol designed for CNV detection using data from high-density SNP genotyping arrays.

SIGNAL PREPROCESSING PROCEDURE FOR SNP GENOTYPING ARRAYS

We have developed the PennCNV algorithm for high-resolution CNV detection from whole-genome SNP genotyping data (Wang et al. 2007); the software is available at <http://www.neurogenome.org/cnv/penncnv>. This algorithm was developed originally using the Illumina HumanHap550 arrays, but can be extended readily to similar SNP genotyping platforms. The PennCNV algorithm uses two measures of signal intensity at each SNP. The Log R Ratio (LRR) is a normalized measure of the total signal intensity for two alleles of the SNP. The B Allele Frequency (BAF) is a normalized measure of the allelic intensity ratio of two alleles. The combination of LRR and BAF can be used to infer copy number changes in the genome (Fig. 1). For example, in the presence of a

FIGURE 1. An illustration of the BAF (upper panel) and LRR (lower panel) values in chromosome 20 of a person with large (>1 Mb) deletion and duplication. The BAF is a measure of the allelic intensity ratio: When a deletion CNV is present, BAF values cluster around 0 or 1 but are absent around 0.5; when duplication is present, BAF values cluster around 0, 0.33, 0.67, and 1, reflecting the AAA, AAB, ABB, and BBB genotypes, respectively. The LRR is a normalized measure of total signal intensity: When a deletion CNV is present, LRR values for SNP markers in this region decrease; when duplication is present, LRR values increase.

deletion, there is a decrease in LRR values and a lack of heterozygotes in BAF values; in the presence of a duplication, there is an increase in LRR values and a splitting of heterozygous genotype clusters into two clusters. To derive the LRR and BAF values, a signal preprocessing procedure is necessary for SNP genotyping platforms (Fig. 2).

The first step of signal preprocessing is the extraction of allele-specific signal intensity values. With the Illumina platform, this step can be performed by the BeadStudio software developed by Illumina as a framework for data analysis and visualization of several types of arrays. The software takes a raw image scan of SNP genotyping arrays and performs a five-step normalization procedure to derive X and Y values for each SNP marker, representing signal intensities for the A and B alleles, respectively. These steps include outlier removal, background estimation, rotational estimation, shear estimation, and scaling estimation (for details, refer to the technical documentation available at <http://www.illumina.com>). The Affymetrix platform uses a set of command line tools (the Affymetrix Power Tools) to perform data normalization and allele-specific signal extraction from raw CEL files generated in genotyping experiments. This step uses a self-normalization algorithm that only requires information contained within the arrays themselves, without relying on any data model built on external reference samples.

The next step in data preprocessing is to derive LRR and BAF from X and Y values for each SNP, based on “canonical genotype clusters” compiled from a set of standard external reference samples (e.g., HapMap samples). The transformation from R and θ into LRR and BAF values adjusts for different chemical characteristics of each SNP, such that values for different SNPs can be compared more readily. This step is designed for and implemented by the Illumina platform (Peiffer et al. 2006), but it can be adapted to the Affymetrix platform as well. Briefly, it involves the calculation of R and θ values, and then the transformation of R and θ into LRR and BAF, respectively.

Although referred to as “polar transformation” (Peiffer et al. 2006), the procedure for R value calculation is not a typical polar transformation. Instead, R is calculated as the total signal intensity of two alleles, or $R = X + Y$. As a normalized measure of total signal intensity, the LRR value for each SNP is then calculated as $LRR = \log_2(R_{\text{observed}}/R_{\text{expected}})$, in which R_{expected} is computed from linear interpolation of canonical genotype clusters (Peiffer et al. 2006) obtained from a set of reference samples. The θ

FIGURE 2. A flowchart of the data processing procedure for deriving the LRR and BAF values used in CNV detection. The Illumina BeadStudio software can be used to calculate LRR and BAF values from raw array image data. For the Affymetrix platform, first extract signal intensity values for all markers using the Affymetrix Power Tools software, and then apply helper scripts for calculating LRR and BAF values. The LRR and BAF values for all markers, together with other relevant information, can then be integrated into an HMM in the PennCNV software for CNV detection.

value measures the relative allelic intensity ratio of two alleles and is calculated as $\theta = \arctan(Y/X)/(\pi/2)$, which ranges from 0 to 1. The BAF refers to a normalized measure of relative signal intensity ratio of the B and A alleles:

$$\text{BAF} = \begin{cases} 0, & \text{if } \theta < \theta_{AA} \\ 0.5(\theta - \theta_{AA})/(\theta_{AB} - \theta_{AA}), & \text{if } \theta_{AA} \leq \theta < \theta_{AB} \\ 0.5 + 0.5(\theta - \theta_{AB})/(\theta_{BB} - \theta_{AB}), & \text{if } \theta_{AB} \leq \theta < \theta_{BB} \\ 1, & \text{if } \theta \geq \theta_{BB} \end{cases}$$

in which θ_{AA} , θ_{AB} , and θ_{BB} are the θ values for three canonical genotype clusters generated from external reference samples (such as HapMap samples).

The transformation procedure described above can be applied to both Illumina and Affymetrix platforms, allowing signal intensity data from these arrays to be analyzed by the PennCNV software. This method has been tested extensively using the Illumina platform for large-scale studies. For example, PennCNV was used for CNV calling in a study that analyzed genome diversity and variations in worldwide human populations by Illumina SNP arrays, and a false positive rate of 0.7% was estimated from duplicated samples (Jakobsson et al. 2008). On the other hand, the signal transformation procedure on the Affymetrix platform has not yet been evaluated as rigorously.

To provide better coverage of known CNV regions and to fill in gaps that are not covered by SNPs, several recently developed SNP genotyping arrays (including the Illumina HumanHap 1M array and the Affymetrix genome-wide 5.0 and 6.0 arrays) contain NP markers on the chip. These markers can be handled in a fashion similar to SNPs for copy number inference, but there are some differences. First, the R -value is calculated as the signal intensity of the NP marker rather than the sum of two alleles. Also, the θ and BAF values cannot be derived for NP markers. Consequently, they are not used in the likelihood calculation. Finally, owing to the use of fewer probes, the variance of LRR values for NP markers may be different from SNP markers. Therefore, the likelihood model parameters for LRR are slightly different between NP markers and SNP markers.

HIDDEN MARKOV MODEL FOR CNV DETECTION

The PennCNV algorithm is an HMM-based algorithm that analyzes signal patterns across the genome and identifies consecutive markers with copy number changes. In the algorithm, the probability of observing a particular copy number state at a particular marker is dependent on the state at the previous marker, so it provides a natural framework for modeling dependence structures between nearby SNPs. As in other studies (Colella et al. 2007), six different possible copy number states are assigned for each marker, representing copy numbers of zero to four, as well as copy-neutral loss of heterozygosity. The signal intensity patterns (LRR and BAF) for five or more copies cannot be readily differentiated from four copies (Wang et al. 2007), so these rare situations are considered to be the same state as four copies.

There are two main components in the HMM for the inference of the most likely path of copy number states: emission probability and transition probability. Emission probability refers to the likelihood of observing the signal intensity patterns (LRR and BAF values) at each marker, whereas transition probability refers to the probability of having copy number changes between adjacent markers. The challenge of the HMM algorithm lies in the inference of the most likely copy number state at each genotyped marker, because this state depends not only on the signal intensity data at this marker, but also on the transition probability from the previous marker.

The emission probability of each marker is a function of LRR and BAF values (except for NP markers), which are assumed to be conditionally independent given the copy number state at that marker in the likelihood calculation. As shown in Figure 1, different copy number states have distinct patterns of LRR and BAF values. The transition probability is dependent on the previous and the current copy number states, as well as the distance between the two markers. Intuitively, copy number states are unlikely to change between nearby markers but more likely to change between sparsely spaced markers (Marioni et al. 2006). The details of the mathematical formula for modeling the likelihood of LRR and BAF are presented in Wang et al. (2007). Briefly, we use the Viterbi algorithm (Viterbi 1967) to infer the most likely states at each SNP efficiently, and then generate CNV calls by identifying stretches of states that are different from the normal state.

FUNCTIONAL ANNOTATION, INTERPRETATION, AND INHERITANCE ANALYSIS OF CNVs

After generating CNV calls using the PennCNV algorithm, one can annotate them functionally and predict their possible effects. CNVs can affect genome function by deleting or duplicating entire genes, parts of genes (such as exons), or important intergenic/intragenic regulatory elements (Feuk et al. 2006). The PennCNV software package contains auxiliary programs to annotate CNVs, such as identifying overlapping or nearby genes and conserved genomic elements for CNV calls. These functional annotations can help formulate new hypotheses on the functional consequences of particular CNVs.

Because the precision of CNV calls is dependent on the technical platform used, their appropriate interpretation requires an understanding of the particular technology used for CNV analysis. Although the “actual CNV” is a stretch of sequence that can start and end at any base pair, the detectable size of a CNV call typically depends on the resolution of the experimental platform. For example, CNV calls could be composed of one or a few bacterial artificial chromosome (BAC) clones (in an array-CGH platform), a stretch of adjacent SNPs (in a high-density genotyping platform), or a stretch of NP markers (in the SNP genotyping arrays with NP markers, or in ultrahigh-density whole genome tiling arrays). Low-resolution CNV calls may cover functionally important genomic elements, but the actual CNVs may not have any functional consequences (Hegele 2007). Likewise, CNV calls from high-density SNP genotyping arrays underestimate slightly the sizes of CNV regions. Furthermore, it is important to recognize that most of these technical platforms can only detect simple duplications or deletions, but are unable to infer more complex structural variations, such as inversions or translocations that can also affect genome function (Tuzun et al. 2005).

Many current CNV studies have examined individuals from the same families (e.g., HapMap families or Centre d’Etude du Polymorphisme Humain [CEPH] families), so Mendelian inheritance can be used to evaluate CNV detection accuracy and generate more confident CNV calls. CNV calls made in both an offspring and his/her parents increase the confidence of the calls, whereas the fraction of CNV calls in the offspring but not detected in the parents (CNV-NDPs) can be used as a composite measure to evaluate CNV calling algorithms (Wang et al. 2007). In addition, family information can also be used directly in CNV calling algorithms for more accurate inference of CNVs. The PennCNV package includes a posterior validation procedure that can incorporate family information to validate CNV calls and resolve boundary discordances between related individuals. Through appropriate prior probability specification, this posterior validation procedure also takes into account possible de novo events, which may be important for human diseases.

SUMMARY

CNVs in the human genome can be detected from high-density SNP genotyping data with well-designed computational algorithms. Further development of methods that better model signal intensity patterns, that handle samples with low-quality signals, and that use more available information, will improve the accuracy of CNV detection and complement current genome-wide association studies in identifying novel disease susceptibility loci.

REFERENCES

- Carter, N.P. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* **39**: **Suppl**: S16–S21.
- Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C., and Ragoussis, J. 2007. QuantiSNP: An Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**: 2013–2025.
- Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**: 75–81.
- Eichler, E.E., Nickerson, D.A., Altshuler, D., Bowcock, A.M., Brooks, L.D., Carter, N.P., Church, D.M., Felsenfeld, A., Guyer, M., Lee, C., et al. 2007. Completing the map of human genetic variation. *Nature* **447**: 161–165.
- Feuk, L., Carson, A.R., and Scherer, S.W. 2006. Structural variation in the human genome. *Nat. Rev. Genet.* **7**: 85–97.
- Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. 2006. Copy number variation: New insights in genome diversity. *Genome Res.* **16**: 949–961.
- Hegele, R.A. 2007. Copy-number variations and human disease. *Am. J. Hum. Genet.* **81**: 414–415.
- lafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**: 949–951.
- Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., and Guerreiro, R., et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998–1003.
- Komura, D., Shen, F., Ishikawa, S., Fitch, K.R., Chen, W., Zhang, J., Liu, G., Ihara, S., Nakamura, H., Hurles, M.E., et al. 2006. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* **16**: 1575–1584.

- Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.
- Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., and Altshuler, D.M., et al. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**: 275–290.
- Marioni, J.C., Thorne, N.P., and Tavaré, S. 2006. BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22**: 1144–1146.
- McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., et al. 2006. Common deletion polymorphisms in the human genome. *Nat. Genet* **38**: 86–92.
- Peiffer, D.A., Le, J.M., Steemers, F.J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C.A., Belmont, J., et al. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16**: 1136–1148.
- Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**: 207–211.
- Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.
- Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.
- Service, R.F. 2006. Gene sequencing. The race for the \$1000 genome. *Science* **311**: 1544–1546.
- Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. 2005. Fine-scale structural variation of the human genome. *Nat. Genet.* **37**: 727–732.
- Viterbi, A.J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **13**: 260–269.
- Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**: 1665–1674.

Cold Spring Harbor Protocols

Copy Number Variation Detection via High-Density SNP Genotyping

Kai Wang and Maja Bucan

Cold Spring Harb Protoc; doi: 10.1101/pdb.top46

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

[Analysis of Microarrays](#) (19 articles)
[Analysis of Microarrays, general](#) (13 articles)
[Bioinformatics/Genomics, general](#) (192 articles)
[Computational Biology](#) (100 articles)
[Genetic Variation](#) (86 articles)
[Genetics, general](#) (374 articles)
[Genome Analysis](#) (191 articles)
[Genomic Analysis](#) (10 articles)
[Microarrays](#) (54 articles)
[Microarrays, general](#) (55 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>