White Paper

Affymetrix® Canary Algorithm Version 1.0

Introduction

Genome-wide association studies seek to identify variation in the human genome which underlies a particular disease, drug response, diagnosis, or prognostic outcome. The cataloging of human variation and subsequent association analysis has traditionally focused on single nucleotide polymorphisms (SNPs). This assessment of common SNP variation in human disease has proven fruitful; more than 50 common variants have been found to be associated with disease such as type 2 diabetes, cardiac, and immunological disease[1, 2].

Recent work, however, has demonstrated that other types of genomic variation—including copy number variants (CNVs)—play a significant role in determining phenotype in common diseases. Copy number polymorphisms (CNPs) comprise a subset of CNVs present in the population. They segregate at an allele frequency greater than 1 percent and are generally smaller than 100 kilobases (kb). Their composition is as much a trait of the human population as it is of individual members, motivating the compilation of CNP maps that catalog CNPs in the human genome.

Like SNPs, CNP alleles (bi-allelic deletion loci have a diploid copy number of zero, one, or two, representing three possible genotypes; bi-allelic duplications have a diploid copy number of two, three, or four) are considered normal without contrary phenotypic evidence. However, unlike typical SNP markers, a CNP is a structural element of the genome and immediately becomes a putative causal variant of its associated phenotype.

In an effort to expand the scope of association studies to include CNPs, Affymetrix, in collaboration with the Broad Institute of Harvard University and the Massachusetts Institute of Technology (MIT), developed Canary, a novel analysis method designed specifically to interrogate CNPs. The Genome-Wide Human SNP Array 6.0 provides Canary with both polymorphic (SNP) and non-polymorphic (copy number) probes for the analysis of CNPs. Additionally, Genotyping Console™ 3.0 provides the industry's only software tool for genotyping both SNPs and CNPs in a genome-wide association study.

The Broad Institute CNP map

To generate the first high-resolution, genome-wide map of common CNPs, the Broad Institute analyzed 263 HapMap samples using the SNP Array 6.0. This analysis yielded a map with accurately defined boundaries and allelic states for each of its CNPs. These CNP regions, referred to here as the Broad Map (called CNV map in Genotyping Console 3.0), have all been visually validated and manually curated to identify their discrete copy number states. The restriction of integral copy number excludes analyses of genotypically heterogeneous samples due to somatically inherited mutations. To be included in the final map, the regions needed to be present more than once in unrelated individuals and demonstrate good predictability for classification. Generally, these regions are smaller than 100 kb.

The Broad Institute implementation of Canary[3,4] is part of a larger package called Birdsuite which, in addition to providing copy number calls in CNPs, also genotypes CNPs in samples holding a copy number of two. Birdsuite is also capable of searching sets of samples for novel CNVs and integrating SNP and copy number calls into a single analysis.

Canary algorithm overview and training

The Canary algorithm, developed by the Broad Institute, comprises an analysis method for assigning a copy number call to the regions in the Broad Map. Using Canary, the physical structure of the genome can be directly associated with phenotype.

At a high level, for each sample at each CNP in the Broad Map, the Canary algorithm computes a single intensity summary statistic using a subset of manually selected probes within the CNP region. These intensity summaries are compared in aggregate across all samples to intensity summaries previously observed in training data to assign a copy number state call. The Broad refers to these previously observed intensity summaries as priors. The priors were generated from 270 HapMap samples processed independently at the Broad Institute and Affymetrix.

The training intensities for each CNP region were clustered and assigned copy number calls of zero, one, two, three, or four. Cluster means, variances, and expected frequencies were tabulated into the prior. All assignments of copy number to clusters computed by the Broad Institute were validated by visual inspection to guard against cases of false clustering

The training set used to generate the cluster information included in the prior file was drawn from 270 HapMap cell lines. This set of 270 cell lines is intended to be large and broadly representative of the human population. It consists of:

- 90 cell lines from the Yoruba people of Ibadan, Nigeria (these are trios, meaning 30 sets of two parents and a child)
- 45 unrelated individuals from the Tokyo area of Japan
- 45 unrelated individuals from Beijing, China
- 30 U.S. trios with northern and western European ancestry

In addition to producing a call for each region, the Canary algorithm also produces a confidence metric that summarizes how well the intensity for the region matches the prior. Confidence is in the range (0, 1) and higher confidence values correspond to better matches.

More importantly, and in contrast to general copy number algorithms, Canary does not assume that a majority of the population has a CNP copy number state equal to two. General copy number algorithms are designed to identify deviations from a normal copy number by comparing each sample to a reference set generated from a large set of individuals. The reference set is used to identify the baseline, which is assumed to represent a copy number state equal to two for phenotypically normal individuals. This is a safe and reasonable assumption to make for large autosomal segments; however, in the case of a specific CNP region, the assumption that a majority of a population has a copy number of two cannot be made with similar confidence. The Canary approach of quantifying prior clustering in training data provides more accurate estimations of copy number for the specific regions defined in the Broad Map.

Figure 1 shows an example of a specific cluster pattern for a region in the map.

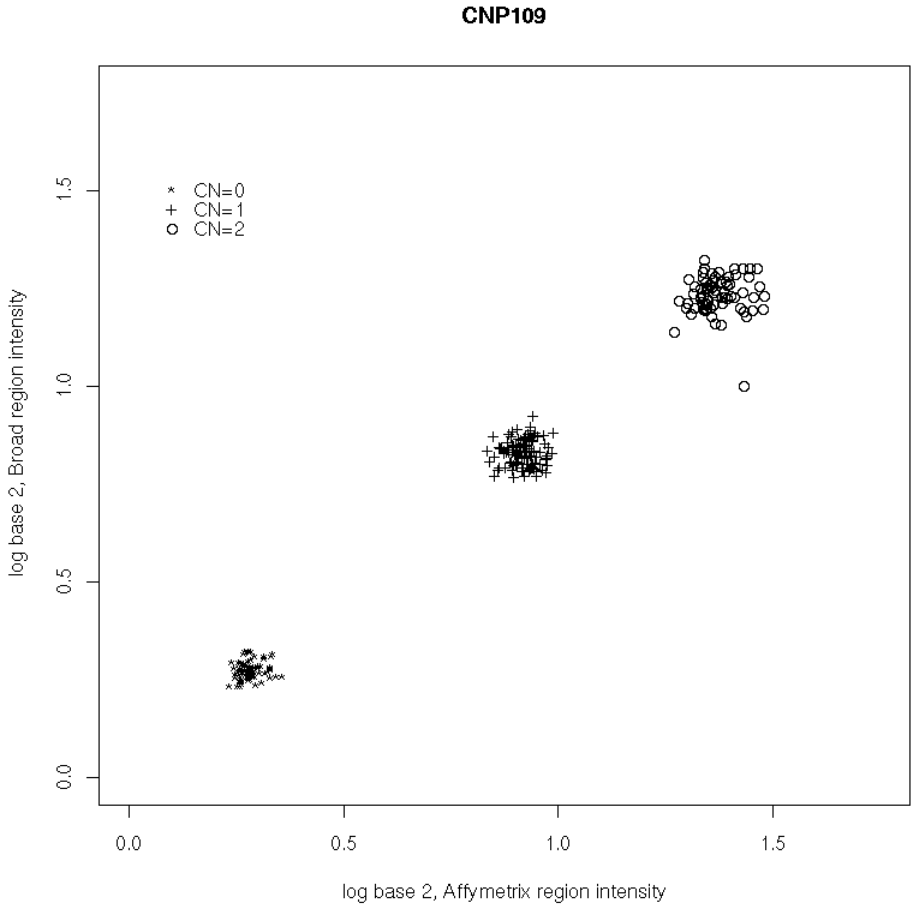

Figure 1: Pairwise replicate sample clustering region CNP109 in the Broad Map illustrates how intensity summaries across samples processed at two different sites cluster similarly.

Canary inputs

The following analysis files are required to run the Canary algorithm: 1) CNP region file; 2) normalization file; and 3) cluster priors. The CNP region file contains CNP boundaries relative to the genome and a list of probes used to calculate the intensity for each region. The normalization file contains a list of probes specified by the Broad Institute for chip-by-chip scalar normalization of probe intensities. The median of these normalization probes provides a multiplicative scalar relative to a targeted intensity. This step of scalar normalization follows quantile normalization performed earlier in the workflow and has only a small effect on the data. The prior file contains the cluster mean, variance, and the expected cluster frequencies. It is important to note that these analysis files are interdependent and their continuity is required for Canary to make accurate CNP calls (i.e., updating one file requires updates to the other two files). Inputs not specific to Canary are the CDF file for mapping CEL file intensities to probes and a set of CEL files containing hybridization data.

Canary outputs

The basic output of Canary consists of a copy number call for each CNP for each sample. The call is reported as an integer directly corresponding to the copy number state of the individual. A call of zero indicates an individual has zero copies of that CNP, a call of one indicates an individual has one copy of that CNP, and likewise for calls up to four. A confidence score for each call is given. The confidence score falls in the range of zero to one, with one indicating a high level of confidence. The confidence score is calculated by dividing the probability of the observation of the copy number call by the sum of probabilities calculated over all possible calls.

Cross-validation

To independently examine the accuracy of the Canary algorithm, CNP calling concordance with quantitative PCR was analyzed. Forty-two CNP regions were chosen at random and analyzed across 30 HapMap samples. Q-PCR appeared to fail in three regions across all samples, and these were removed from the analysis. The observed concordance between the Broad assigned genotypes in the training data and Q-PCR across all HapMap samples and CNP regions analyzed was 96 percent.

Reproducibility

To characterize the performance of the Canary algorithm and the relationship between the confidence score and the reliability of a CNP call, the following experiment was performed: 567 HapMap samples were processed at seven independent sites, analyzed with Canary, and the results compared to the CNP call curated by the Broad Institute.

The CNP calls across 567 samples were ranked by confidence score. The CNP calls with the lowest confidence scores are most likely to be unreliable, and thus were considered as top candidates for "no-calls" (the idea being that the limitations of Canary can be addressed and at the same time the overall quality of remaining data summaries improved by ignoring no-calls). The proportion of this excluded data in a data set is referred to as the "no-call rate." The complement of no-call rate provides the call rate. For example, to evaluate concordance for a call rate of 90 percent, exclude the data corresponding to the lowest 10 percent of confidences.

In Figure 2, the no-call rate is plotted on the horizontal axis and concordance is plotted on the vertical axis. Without considering confidence—that is, without discarding any data—the overall concordance between observed CNP calls and the curated CNP calls is 97 percent. The maximum concordance that can be achieved without dropping all but a few observations is 99 percent.

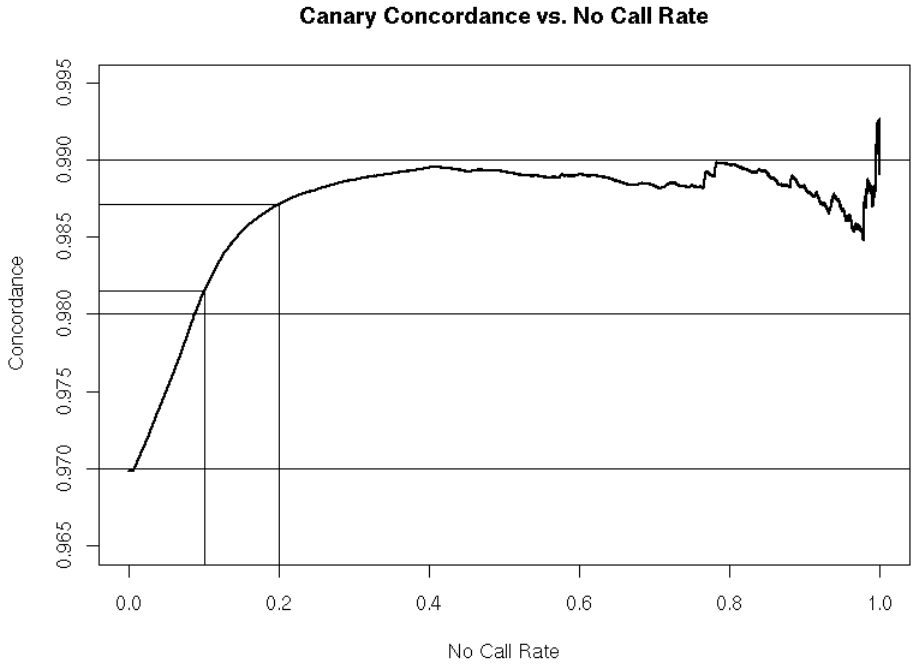

Figure 2: Concordance with respect to no-call rate.

If a no-call rate of 10 percent is tolerated, concordance exceeds 98 percent, which is more than half of what can be achieved using confidence as a filter. At a no-call rate of 20 percent, the slope of the concordance curve indicates that the return of discarding data with low-ranking confidence scores

has substantially diminished. The jagged feature of the concordance curve at the right is due to computing a fraction of concordant CNP calls over a decreasing number of observations. In Figure 3, the upper curve shows actual confidence scores in comparison to their rank. The lower curve shows the fraction of incorrect calls accounted for by the no-call rate. We see here that confidence initially rises rapidly as rank increases. Similarly, the proportion of incorrect calls is rapidly accounted for. Once a no-call rate of 0.2 is tolerated, gains in confidence have almost completely tapered off. At the same time the gain in incorrect calls is approaching a linear increase toward one. This suggests that past a no-call rate of 0.2, the confidence score contains no further information in regard to concordance.

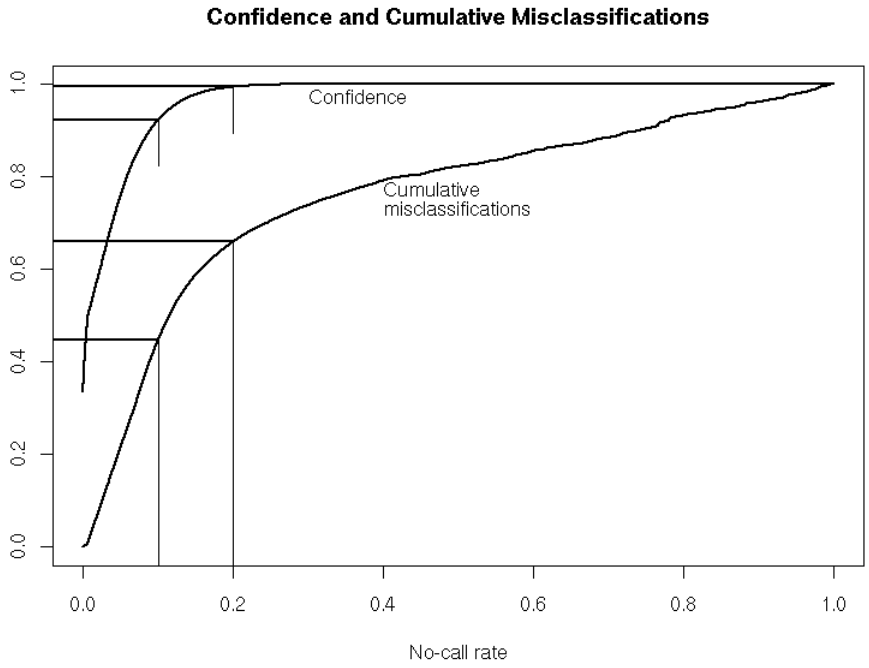

Figure 3: Criteria for establishing no call rates. Observations are ranked by confidence score. Top curve is concordance versus rank; lower curve shows accumulation of total incorrect calls versus rank.

The following recommendations are based on this analysis:

1. It is recommended that you filter CNP calls based on confidence score to reduce the number of misclassifications.

2. It is recommended that you use a no-call rate of 10 percent to filter data. This results in decreasing the number of misclassifications by 45 percent, at the expense of only 10 percent of the data.

3. The actual no-call rate used for an analysis needs to be set by the individual investigator, based on the desired false-positive/false-negative comfort levels.

MAPD quality control

The median absolute pairwise difference (MAPD) score is a standard Affymetrix copy number quality control indicator that increases as data quality becomes more suspect. The indicator is computed on a per-sample basis. In Figure 4, the relationship between MAPD score and the number of discordant calls indicates that Canary, like any algorithm, produces fewer incorrect calls results as data quality improves.

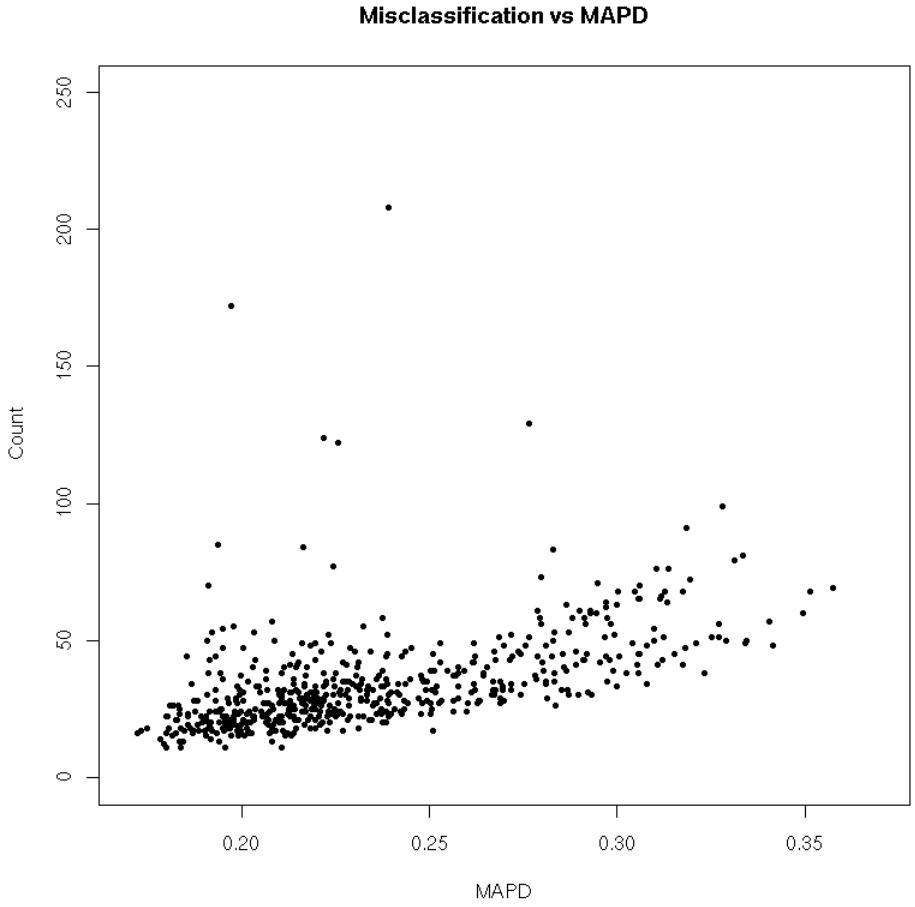

Figure 4: Plot of incorrect calls per sample versus MAPD score. Six outliers observed need more investigation and are not shown.

Conclusions

For Canary CNP analysis, what are the recommended metrics for sample QC and for filtering CNP region calls?

For sample QC in Canary analysis, the same metrics for genotyping and copy number QC can be applied. You should verify that samples pass one or both of these metrics to anticipate good results using the Canary algorithm. Therefore, samples with a Contrast QC metric greater than 0.4 should be included in the CNP analysis data set (like genotyping QC). If you have run the CN5 copy number algorithm in Genotyping Console 2.1, samples with a MAPD metric less than or equal to 0.4 should be included in the CNP analysis data set.

Additionally, filtering on CNP call confidence to reach a no-call rate of 10 percent across the entire data set is recommended. This is based on internal concordance data which indicated that a 10 percent no-call rate eliminated 50 percent of CNP region misclassifications (Figure 3). The CNP call confidence threshold to achieve a 10 percent no-call rate will vary slightly for each data set. For internal data, the confidence threshold was 0.92, and a similar threshold should be appropriate for most data sets. After applying the confidence filter in third-party analysis software, you should verify that the no-call rate is approximately 10 percent.

Lastly, the expected frequency distribution for a call across multiple samples is used in the assignment of the final call; therefore, a minimum of 30 samples should be processed as a batch.

References

1. Bowcock, A. M., *et al.* Genomics: guilt by association. *Nature* **447**:645-646 (2007).

2. Altshuler, D. and Daly, M. Guilt beyond a reasonable doubt. *Nature Genetics* **39**:813-815 (2007).

3. McCarroll, S. A., *et al*. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* DOI: 10.1038/ng.238 (Advance online publication, 2008).

4. Korn, J. M., *et al*. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* DOI: 10.1038/ng.237 (Advance online publication, 2008).