Application Note: Informatics

A Patient-Centric SNV-CNV Pipeline

Methodology for estimating DNA copy-number changes from microarray data.

Introduction

This application note outlines the Illumina methodology for estimating DNA copy number for data produced on Affymetrix Genome Wide Human single nucleotide polymorphism (SNP) 5.0 and 6.0 arrays on the BaseSpace Correlation Engine. Within a patient-centric context, data are obtained for an individual patient rather than a batch. Also, patient data are often supplied without a matching reference. The copy-number variations (CNVs) and copy-number aberrations (CNAs) identification pipeline must account for both these factors.

Methods and Results

Data Normalization

When processing array data, it is essential to apply normalization methods that reveal biological variation by correcting for unwanted technical variation. Each platform tends to have specific biases that require specific normalization approaches.

Affymetrix SNP arrays were originally developed for genotyping in genome-wide association studies (GWAS). In this context, genotype calls are estimated independently for each single-nucleotide variant (SNV) by comparing signal distributions across samples. As a result, a handful of methods exist to accomplish across-sample normalization.

Copy number studies also need the SNV signals to be comparable across samples. Copy-number changes are determined by comparing a sample against a reference, or groups against other groups; therefore, the underlying signals must be similar. The same normalization approaches used for GWAS are also used for copy-number studies.

In addition to across-sample normalization, true copy-number signals are expected to be locally constant along the genome; ie, more often than not, neighboring probes should have identical copy number. Many downstream segmentation algorithms for detecting copy number changes exploit this property. Therefore, it is essential to make CNV/SNV signals comparable across loci within a sample, and across samples. Methods used to perform within-sample normalization require a reference group. The sample and reference data must be comparable; therefore, across-sample normalization is performed before within-sample normalization.

Across-Sample Normalization

The Illumina BaseSpace Correlation Engine takes advantage of the aroma Affymetrix software package[1] and, in particular, the normalization methods provided in CRMA Version 2,[2] to address the technical bias present in Affymetrix 5.0 and 6.0 arrays. CRMA v2 is a software suite designed to allow single-array processing to handle individual patient data.

The methods in CRMA v2 address biases due to allelic-crosstalk between probes measuring the A allele and the B allele. They account for probe-sequence effects present in arrays that have small probe sizes and adjust for fragment-length effects produced from the 2 different enzymes used on the array. They also perform probe-level summarization, converting the A-allele signal and B-allele signal into a total copy-number signal for each locus. These normalization steps improve the comparability of samples processed in different environments.

Within-Sample Normalization

The signal intensity at a locus is presumed to be proportional to the copy number at that locus. However, the proportionality of signal-to-copy number is locus-specific, due to cross-hybridization, probe thermodynamics, and other factors. These local affinities must be removed to make copy-number signals across neighboring loci meaningful. The locus affinities are not sample-specific and, therefore, can be effectively minimized by dividing the sample locus signal with the same locus from a reference group signal.

The choice of reference group requires careful consideration. How a reference is selected can significantly impact patient copy-number results.

The Reference Group

The selection of the reference group used for within-sample normalization and downstream analysis can noticeably affect the outcome of the patient copy-number results. In particular, depending on the goal of the analysis, the choice of reference group can be important for producing relevant copy-number data.

Group Reference Case Study I

Illumina performed extensive internal analyses to understand the effects of reference group selection. The results of one such analysis are highlighted here, using The Cancer Genome Atlas (TCGA) Glioblastoma Multiform data.[3,4] These data have been carefully curated. Two reference groups were created: one containing 366 patients with DNA taken from "normal blood" and another containing 25 patients with DNA taken from "normal adjacent tissue."

One normal blood sample and one "primary tumor" sample were randomly selected to use as a hypothetical patient. Across-sample normalization was performed for all data sets to remove unwanted technical variation. This was followed by within-sample normalization for each hypothetical patient for 2 scenarios: one when the 366 normal blood samples were used as a reference group, and another when the 25 normal adjacent tissue samples were used as a reference group.

After normalization, segmentation was performed to find genomic regions that indicated a copy-number change between the patient and the reference group. The regions of CNAs reported for each hypothetical patient were consistent, regardless of which reference group was used. This result, while encouraging, was expected, because the tested reference groups came from the same study as the hypothetical patients. When possible, it is best to use reference samples from the same study.

Group Reference Case Study II

Often, the patient to be analyzed will not have accompanying samples from which a reference group can be constructed. To test such situations, Illumina took the 2 hypothetical patients from Case Study I, and used a publicly available "normal brain tissue" data set, GSE9635, downloaded from the Gene Expression Omnibus (GEO).[5]

This cohort of 30 patients was normalized using CRMA v2 and grouped as a reference. The probe-signal variance across these 30 patients was 20% higher than the probe-signal variance for the 25 normal adjacent tissue TCGA samples used in Case Study I. Next, within-sample normalization, and segmentation for each hypothetical patient was performed, using the GEO cohort as a reference. Over 50% of the regions reported when the GEO samples were used as a reference were also reported when the TCGA samples were used as a reference.

This analysis demonstrated 2 results. First, regions with large copy-number changes were overwhelmingly present in both result sets, suggesting that strong signals were detectable regardless of the reference group used. Second, regions with moderate to low signal intensities in one result set were more likely to be missing in the other.

Paired Reference Case Study

For the TCGA data summarized in Case Study I, many of the patients have both a solid tumor data set and a normal blood data set, permitting a paired analysis. This analysis was performed by using the solid primary tumor sample as the hypothetical patient and the patient's normal blood sample as the reference.

Comparing the copy-number changes for this paired approach with the copy-number changes reported from Case Study I (where the 366 normal blood samples were used as reference), a distinct difference in result sets was evident. Regions that were present when using the 366 patients as a reference, but not present in the paired approach, were overwhelmingly regions known to have common copy-number variations. This suggests the presence of germline CNVs in the patient, which is negated when the analysis uses the patient's paired normal data set as reference. Therefore, if the researcher wishes to capture purely somatic copy-number changes, a paired approach is ideal. For cases when a paired data set is unavailable, removing regions known to have common CNVs from the result set should be an available option.

Germline CNVs vs. Somatic CNVs

As described in the paired reference case study, using a group as a reference permits the patient's germline CNVs to be reported in the result set. Often, these germline variations are common CNVs and, therefore, should be removed from the results when somatic mutations are of primary interest. However, special care must be taken when removing common CNVs, because what is considered common is often population-specific.

Segmentation

Illumina used the algorithm Circular Binary Segmentation[6] to segment the normalized data. A key parameter for this algorithm is the number of probes required within a region to support the assertion that a copy-number change exists. The convention is to require between 3 and 5 probes. Illumina internal analysis suggests a higher probe count may be desired to limit false-positive aberration calls. Illumina implemented this change to set the minimum number of probes per region to 9, reducing noise as a result.

The analysis pipeline will often be processing individual patient data. As a result, the signal for each probe is a single measurement from the underlying probe distribution and, thus, can demonstrate significant variation. Typical copy-number studies have several samples in each group, allowing for a more robust estimate of each probe signal. Single-patient studies, however, contain sufficient variation that the probability of 5 neighboring probes having a significant signal by chance is high.

Copy-Number Calling

The process of assigning a copy-number count to regions that are significantly changed between the patient and reference group has been an area of intense research. As yet, no method has gained universal acceptance. Often researchers simply assign labels (eg, "Gain" and "Loss") to regions whose signal is above some fixed threshold. Illumina adopted this approach as well, with a slight modification. The thresholds used for calling a copy-number change are a function of the underlying biology and sample variation. Internal investigation showed that having a fixed threshold can be too general, given the inherent variation in single-patient study designs.

The software suite CRMA v2 reports estimated copy-number values for segmented regions. The BaseSpace Engine maps these estimates to discrete copy-number counts based on the determined thresholds. The mapping assigns regions a copy-number count label (-2, -1, 0, 1, 2,…). The label also provides a level of abstraction that permits copy-number regions to be comparable across patient/study environments in a metaanalysis setting.

Removing Common CNVs

Using a global aggregate reference group when analyzing patient copy-number data will result in the patient's germline CNVs being present in the result set. Often, the primary interest is finding somatic copy-number changes, rather than germline variations. Germline variations tend to be common among a large set of people. Therefore, the ability to remove common CNVs (eg, those reported in the Database of Genomic Variants)[7] is desirable. However, care must be taken when removing common variations using a generic repository, because what is considered common is often population-specific.

Pipeline Workflow

The basic workflow of the BaseSpace Engine SNV-CNV analysis is illustrated in Figure 1.

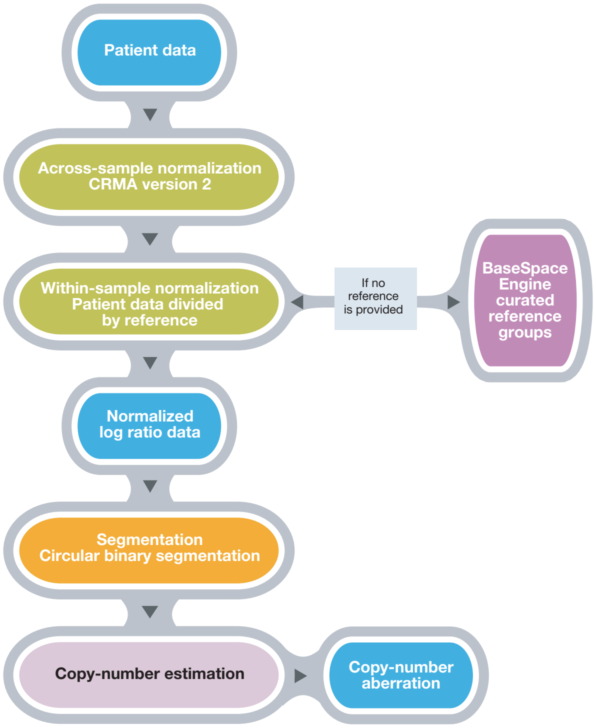
Patient data
Across-sample normalization CRMA version 2
Within-sample normalization Patient data divided by reference
If no reference is provided
BaseSpace Engine curated reference groups
Normalized log ratio data
Segmentation Circular binary segmentation
Copy-number estimation
Copy-number aberration

Figure 1: BaseSpace Correlation Engine Data Processing Pipeline

Conclusions

A growing body of evidence links CNAs with disease. Thus, an analysis pipeline capable of processing individual patients—and comparing the results across patient cohorts—is highly desirable. While the current copy-number analysis pipeline takes into account the many complexities associated with individual patient analysis, potential improvements are in development.

The BaseSpace Correlation Engine has an unprecedented amount of curated patient-level data that can be used to create customizable reference groups for processing copy-number studies. Selecting an appropriate reference group is crucial for producing quality result sets. As the number of inhouse patient data sets increases, so does the catalog of curated reference data sets. This growing resource will greatly improve the BaseSpace Engine processing capabilities. Also, having many data sets allows Illumina to test methodologies across several environments to make sure that the BaseSpace Engine CNV/CNA pipeline can handle different analysis settings.

References

1. www.aroma-project.org
2. Bengtsson H, Wirapati P, Speed TP. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*. 2009;25(17):2149–2156.
3. American Cancer Society: Cancer Facts and Figures 2010. Atlanta, GA.
4. Levin VA, Leibel SA, and Gutin PH. Neoplasms of the central nervous system. In: DeVita VT Jr, Hellman S, Rosenberg SA, editors. Cancer: Principles and Practice of Oncology (6th ed.). 2001;2100–2160.
5. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci USA*. 2007;104(50):20007–20012.
6. cran.r-project.org/web/packages/PSCBS/PSCBS.pdf
7. dgv.tcag.ca/dgv/app/home

For Research Use Only. Not for use in diagnostic procedures.

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.