>

Researchers Find Arrays Miss Many CNVs; Recommend Combining Platforms, Algorithms to Improve Results

June 21, 2011

Researchers Find Arrays Miss Many CNVs; Recommend Combining Platforms, Algorithms to Improve Results

By Justin Petrone

An international team of scientists recently completed a "massive initiative" to evaluate the ability of available microarray platforms and accompanying algorithms to discover copy number variation in genes implicated in diseases.

Led by researchers at Uppsala University in Sweden and the Hospital for Sick Children in Toronto, the team tested 11 different microarray platforms and found that current methods of analysis only detect a portion of the CNVs in the genome in any given experiment.

Other institutions contributing to the study included Brigham and Women's Hospital in Boston and the Wellcome Trust Sanger Institute in Hinxton, UK. A paper detailing the work was published this month in *Nature Biotechnology*.

According to the paper, for most microarray platforms, reproducibility of replicate samples is below 70 percent and results from different analysis tools applied to the same data typically show less than 50 percent reproducibility. These and other findings led the researchers to conclude that using multiple microarray platforms and a combination of algorithms will lead to a better discovery yield of CNVs.

"The results of different platforms run on the same DNA give surprisingly different results," Lars Feuk, an Uppsala University researcher and corresponding author on the paper, told *BioArray News* this week.

According to Feuk, the researchers also found that for smaller variants, or variants in complex regions, the reproducibility between replicate experiments for the same sample with the same platform was "not as high as we would have expected."

Additionally, Feuk said the team was "surprised" by the "vast differences between various calling algorithms applied to the same raw data, and the low overlap between them."

He said that the number of true positives in a study can be "significantly increased by using a combination of algorithms" to analyze the data.

"There are many genome-wide CNV studies whose design and analysis plan will be much better informed by the results and standardized datasets generated by this massive initiative," Steve Scherer, director of the Centre for Applied Genomics at Sick Kids, told *BioArray News* this week.

The authors of the paper argue that despite their findings, current methods can detect large CNVs for clinical diagnostic purposes because "large CNVs with poor reproducibility are found primarily in complex genomic regions and would typically be removed by standard clinical data curation."

They also warned that the "striking differences" between CNV calls from different platforms and analytic tools call for "careful assessment of experimental design in discovery and association studies" and "strict data curation and filtering" in diagnostics.

The research team has released the raw data from the study — comprising more than 180 independent data sets — through the Gene Expression Omnibus database in hopes that the data will enable other researchers to independently evaluate their data and benchmark new algorithms.

"We expect that the use of microarrays will continue to grow over the next few years and that they will be a mainstay in genome-wide diagnostics for some time," the authors wrote in the paper. "By making these data sets available to the research community, we anticipate that they will be a valuable resource for further analyses and development of CNV-calling algorithms and as test data for comparison with additional current and future platforms."

CGH and SNP Arrays

As part of the study, the authors evaluated of 11 microarrays commonly used for CNV analysis in an attempt to understand the advantages and limitations of each platform for detecting CNVs. Six control samples were tested in triplicate on each array and each data set was analyzed with one to five analytic tools, including those recommended by each array producer. This resulted in more than 30 independent data sets for each sample, which they compared and analyzed.

The researchers broke the 11 arrays into three categories: comparative genomic hybridization arrays, SNP arrays, and those containing SNP and CNV probes. Specifically, they ran their samples on the Wellcome Trust Sanger Institute's Whole Genome Tiling Path Array; Agilent Technologies' one and two-sample 244K arrays; Roche NimbleGen's 720K and 2.1M CGH arrays; Affymetrix's 500K and SNP 6.0 Arrays; and Illumina's 650Y, 660W, 1M, and Omni BeadChips.

[pagebreak]

According to the authors, the newer arrays, which included subsets of probes specifically targeting CNVs, outperformed older arrays both in terms of the number of calls and the reproducibility of those calls. Analysis of the deviation in breakpoint estimates based on the number of probes showed that this difference was not only due to increased resolution, but was also consistent with improved individual probe performance in the newer arrays. "These results highlight that newer arrays provide more accurate data, whether the focus is on smaller or larger variants," the authors concluded.

They also argued that for large cohort studies, it is important that all experiments are processed in one facility. "Even though the data from different sites can be quite similar, they still differ enough to create problems in association analyses," the authors wrote. "It is also clear that comparison of data sets resulting from different platforms and different analytic tools will cause problems in association analyses and may create false association signals," they noted.

While the authors seemed to agree that using multiple, newer arrays was a better strategy for CNV discovery, they did not urge researchers to chose CGH platforms over SNP arrays or vice versa.

"In my view, the best choice of array or array type really depends on your study design and what you are looking for," Feuk said. "In many applications there is a need to have access to SNP data in addition to CNV data," he said. "If you are only interested in CNVs, then a CGH array may be a better approach."

According to Feuk, CGH arrays "generally have better probe coverage in complex and duplicated regions of the genome." Ideally, he said that researchers should also look at probe distribution before starting a project to ensure that the array covers regions that may be of specific interest to the researchers' application or research question.

Feuk noted that his lab is currently running two array-based projects that require both SNP and CNV information, and that he is using Affy SNP 6.0 in one and the Illumina Omni chips in the other.

While the authors of the recent paper described their evaluation as "exhaustive," they did not test some of the new platforms that are in development that contain both probes for CGH as well as SNP content. Agilent launched such a product last year while Roche NimbleGen plans to soon follow suit (*BAN 9/14/2011*, *BAN 3/29/2011*).

Feuk said that it is "great that the array producers are thinking about ways to combine CGH and SNP typing." He said that even if only CNVs are of interest in a study, the presence of SNPs can add further support to the CNV calls. SNPs also work for identification of loss-of-heterozygosity regions and provide the ability to track inheritance through families better than CNV data alone, he pointed out.

"I still think there is a lot of room both for SNP arrays that can be used for CNV analysis, and for CGH arrays that also provide some SNP data," said Feuk. "My recommendation would be that the CGH arrays are complemented with a SNP-calling chemistry that allows actual genotyping of individual markers, rather than being able to detect LOH over large regions by combining information from several probes."

Calling Algorithms

As part of the study, the authors investigated the effects of using different CNV-calling algorithms and determined that the choice of analysis tool can be as important as the choice of array for accurate CNV detection.

"Different algorithms give substantially different quantity and quality of CNV calls, even when identical raw data are used as the input," they wrote. This finding has "important implications both for CNV-based, genome-wide association studies and for the genetic diagnostics field," they added.

Specifically, they showed that algorithms developed for a certain data type, such as Birdsuite for the SNP 6.0 and DNA Analytics for Agilent data, "generally perform better than platform-independent algorithms or tools that have been re-adapted for newer versions of an array."

Feuk said that his main recommendation to researchers is not settle for a single analysis tool. "Try a few different approaches and see what works well for your data, and ideally, use several tools in parallel to get a better sense for the validity of different CNV regions," Feuk advised.

He also stressed that it is "important" to perform proper quality control of the data. "We even find that algorithms are differently suited for noisy and clean data from the same platform," said Feuk. "It is good to know where your data falls in terms of quality, and to use different algorithms."

Ultimately, using multiple algorithms minimizes the number of false discoveries, allowing for greater experimental validation by qPCR, which is typically greater than 95 percent for variants larger than 30 kilobases, the authors concluded in the paper. "Because the algorithms use different strategies for CNV calling, their strengths can be leveraged to ensure maximum specificity."

Have topics you'd like to see covered in *BioArray News*? Contact the editor at jpetrone [at] genomeweb [.] com

Related Stories

- Affymetrix to Launch Next-Gen Cytogenetics Array in Q3, Plans FDA Submission
  May 10, 2011 / BioArray News
- Roche NimbleGen to Expand CGH Array Menu, Plans Additional High-Density Chips
  March 29, 2011 / BioArray News
- Teams ID Genetic Glitches in Induced Pluripotent Stem Cells
  March 2, 2011 / GenomeWeb Daily News
- For Cytogenetic-Test Providers, the Devil's in the Interpretation of Array Results
  July 6, 2010 / BioArray News
- Facing Prospect of Increased FDA Regulation, Vendors Look to Clear Cyto Chips for Clinical Use
  April 13, 2010 / BioArray News