Drupal.behaviors.print = function(context) {window.print();window.close();}>

genomeweb

UW Software Supports Analysis of Rare and Common Copy Number Changes Using Exome Sequence Data

June 12, 2012

UW Software Supports Analysis of Rare and Common Copy Number Changes Using Exome Sequence Data

By Andrea Anderson

A University of Washington-led team has developed a computational method for detecting and genotyping copy number variations based on read depth patterns in exome sequence data.

The developers have released an open source version of the software, known as Copy Number Inference From Exome Reads, or CoNIFER.

"In light of the tens of thousands of exomes anticipated to be sequenced and analyzed in the near future, we believe this method will have widespread application for the discovery and association of both rare and common copy number variation in disease," University of Washington genome sciences researcher Evan Eichler, the study's corresponding author, and colleagues argued, "and will complement existing methods to discover single-nucleotide variation from exome-sequencing data."

As the researchers reported in a recent study published online in *Genome Research*, the approach is proving useful not only for finding rare CNVs in coding sequences, but also for genotyping more commonly occurring copy number polymorphisms, such as sites containing genes found at variable copy numbers from one individual to the next or between different human populations.

"What we found when we developed CoNIFER was that not only were we able to find rare CNVs — such as the deletions and duplications that have been implicated in the pathogenesis of autism and other diseases — but that we were also able to characterize CNPs and specifically estimate the copy number of some of these CNPs," the study's first author, Niklas Krumm, a graduate student in Eichler's University of Washington lab, told *In Sequence*.

Within hundreds of exomes assessed in their study, the researchers found that their algorithm detected nearly 80 percent of the rare CNVs identified by SNP array-based genotyping, as well as additional CNVs and CNPs smaller than the detection limit of existing array-based CNV detection methods.

The method is not intended to entirely replace array CGH or SNP-based methods for characterizing copy number variants — particularly since it relies on exome sequence reads that represent just a fraction of the genome. But results so far suggest that it can augment existing methods and provide a more detailed look at genic regions.

"On some level, it's just an added benefit," Krumm said. "You basically get these things for free if you've done the exome sequencing."

"You are going to get 80 percent of the genic events that you would find using a high-density array CGH platform using just CoNIFER — and you're going to find a couple more events that you wouldn't have found otherwise," he explained.

Because it uncovers smaller-than-usual CNVs and can be used to genotype CNPs found at fairly high copy number, those involved in the study argued that CoNIFER should complement other detection methods, both as a means of characterizing normal variation in gene-coding stretches of the genome and as a strategy for finding new disease associations.

For their part, for instance, the team plans to use the computational approach to help track down CNVs and CNPs in the thousands of exomes sequenced at the University of Washington as part of the National Heart, Lung, and Blood Institute Exome Sequencing Project.

Although exome sequencing has taken off as a means of finding variants such as SNPs and small insertions and deletions in gene-coding portions of the genome, identifying larger deletions and duplications in the exome has generally been more challenging, Krumm explained.

While methods have been developed to discern CNV patterns from whole-genome sequencing data, peculiarities in exome sequence data — such as non-uniform sequence capture or coverage — threw a wrench in efforts to plug exome sequence information into the same tools.

So, too, did other types of so-called batch effects — unpredictable and difficult-to-track errors that can stem from everything from sample preparation variability to slight differences in the chemistry and can lead to false-positive CNV and CNP calls.

Likewise, methods such ExomeCNV that have been developed for finding copy number alterations in cancer exomes using sequence data from matched tumor-normal pairs fell short for finding CNVs in exomes from non-cancerous samples.

While Krumm called ExomeCNV a "simple and effective tool" for finding CNVs from matched tumor and control exome data, he explained that it is less effective when looking at normal genomes, where batch effects interfere with authentic CNV signals.

As described in the *Genome Research* study, the University of Washington and University of Virginia group found that ExomeCNV yielded a slew of false-positive copy number changes in situations where the researchers were not looking at well-matched disease and control exomes that had been prepared and sequenced the same way and under comparable conditions.

Their solution relies both on a normalization method called RPKM — which takes into account the size of a given exon, the number of reads landing on it, and the total number of reads — and on another method known as singular value decomposition, or SVD, which uses a statistical approach to identify the largest contributors to variance in the data.

By combining the two methods, Krumm explained, the researchers could take advantage of the read depth information available in the exome sequence data while minimizing batch effect-related false-positive copy number predictions.

"[Batch effects] are kind of difficult to track and almost impossible to correct factors," Krumm explained. "But when you use a statistical approach and when you combine data from multiple exomes together, you can start to parse out what is actually a batch effect that affects most or all of the samples versus … actual biological signals found within the exome data."

Initially, the researchers came up with the computational methods as a means of screening their exome sequences for very large CNVs, Krumm explained, "to make sure we weren't missing anything crazy."

But the team soon found that the method had quite a bit of power for finding not only large copy number changes but also much smaller polymorphisms in exomes, including some alterations that are smaller than the detection limit of arrays that span the entire genome.

For instance, after taking a crack at using the method to test eight HapMap exomes with well-characterized copy number profiles, the researchers turned their attention to 366 exomes representing 122 parent-child trios sequenced as part of a study on autism spectrum disorder.

Within the 109 ASD trios for whom corresponding Illumina 1M SNP array-based CNV data was available, the researchers found 317 apparent CNVs affecting at least three exons.

Once they tossed out potential variants found in segmentally duplicated regions, somatically rearranged loci, and pseudogenes, the investigators were left with 124 candidate copy number variants that included eight candidate *de novo* CNVs, 87 inherited CNVs, and 29 possible CNPs.

Of these, the team verified 77 percent using SNP array data. Additional testing, including quantitative PCR or array CGH with custom arrays, eventually helped to confirm even more of the CoNIFER-predicted CNVs, leading to a precision rate of 94 percent for the computational method in the ASD trio exomes.

In particular, the researchers reported that the algorithm edged out array-based methods for finding CNVs that were around 14,000 bases long. It also appears well suited for determining copy number at CNPs present in up to eight copies.

Conversely, the CoNIFER software picked up 76 percent of 109 CNVs previously identified in the ASD exomes using SNP arrays.

"If you have exome data and you have array data, combining the two really lets you discover all the genic events — or a much greater fraction of the genic events in your samples," Krumm explained.

Because CoNIFER compares exome sequences to one another to help distinguish authentic copy number patterns from experimental artifacts, it hinges on the availability of data from multiple exomes.

Krumm recommends including information on at least eight exomes at once, though he explained that the minimum number of exomes needed may be as high as 20 depending on the consistency of the exome sequence data being considered and the methods used to generate it.

For labs that want to compare their exomes against more sequences than they or their collaborators have generated, he pointed to the possibility of tapping into publicly available exome data when applying the CoNIFER pipeline.

Again, though, he noted that the quality of the results may vary somewhat depending on how closely the experimental procedures used to generate these public sequences gibe with those used to generate the exomes of interest.

Still, while exomes considered in the current study were all sequenced using either the Illumina HiSeq 2000 or the GAII, Krumm said that CoNIFER appears to be "very robust to different capture technologies and different sequencing technologies."

Among the exomes included in the current study, for instance, were exomes that had been captured using the Agilent SureSelect exome capture kit, two versions of the Roche NimbleGen SeqCap EZ capture kit, and other NimbleGen and Agilent systems.

"We realize that exome capture technology is constantly evolving and that [companies are] always changing the targets and things like that," he said. "We've successfully mixed experiments using different capture [methods], from different sequencing runs, from different sequencers, different read lengths."

Even so, the study authors noted that "care should be taken when combining data across significantly different platforms — in these cases, only the common set of probes between platforms should be used in order to avoid false negatives."

The study authors suggest using at least 50 million on-target reads per exome to help maintain a sufficient signal-to-noise ratio.

They also noted that the approach appears to be work better when paired with short-read alignment tools that allow the same read to be mapped multiple times and cautioned that it does not appear to be particularly compatible with exomes generated using whole-genome amplified DNA, which interferes with the analysis.

Andrea Anderson is a senior science reporter for GenomeWeb Daily News, covering genomics research studies and translational research. E-mail her here or follow GWDN's headlines at @DailyNewsGW.

Related Stories

- In Print: Last Week's Sequencing-Related Papers of Note
  June 12, 2012 / In Sequence
- In Print: Sequencing-Related Papers of Note of the Last Two Weeks
  June 5, 2012 / In Sequence
- In Print: Last Week's Sequencing-Related Papers of Note
  May 15, 2012 / In Sequence
- Virginia Tech to Lead NSF-funded Amphibian Microbe Study
  November 8, 2011 / GenomeWeb Daily News
- Statistical Strategy Gleans Historical Population Patterns from Individual Genome Sequences

July 13, 2011 / GenomeWeb Daily News