Accurate Detection of Structural Genetic Changes Using NGS Data to Reduce the Proportion of Unsolved Cases at Low Resource Cost

Sri N. Shekar[2], Lisa Herta[2], William J. Salerno[1], Adam C.English[1,2], Catherine A. Brownstein[4,5], Joseph Gonzalez-Heydrich[4], Adina Mangubat[2], Jeremy Bruestle[2], Eric Boerwinkle[1,3], Richard A. Gibbs[1]
1. Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA; 2. Fabric Genomics™, Inc., Oakland, USA; 3. Human Genetics Center, University of Texas Health Science Center, Houston, Texas, USA;
4. Boston Childrens Hospital, Boston, USA; 5. The Manton Center for Orphan Disease Research, Division of Genetics and Genomics, Hospital, Boston, USA. nshekar@fabricgenomics.com

BACKGROUND

Could de novo structural variants be associated with rare disorders?

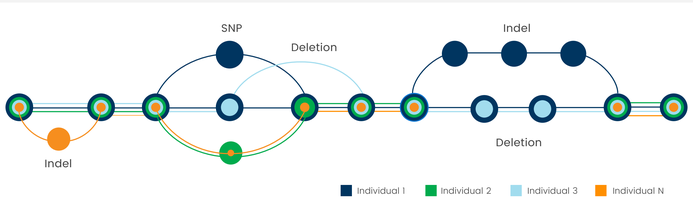
Next generation sequencing and associated bioinformatics has made it possible to detect SNVs and small indels. Callers for SNVs and small indels, using the method of aligning reads, are able to detect SNVs and small indels (<30bp) with high sensitivity and a low false discovery rate. As such, it has been possible to identify those SNVs and indels that arise de novo. Consequently, this has reduced the number of unsolved cases of rare disorders. However, there remain a number of these unsolved cases. Although the number of de novo indels (>30bp) and structural variants may be far fewer than the number of de novo SNVs (Acuna-Hidalgo et al., 2016), it is possible that accurate detection of these variants may further reduce the proportion of unsolved cases.

Currently, it is difficult to detect de novo structural variants

Methods to detect structural variants often rely on heuristics related to whether reads are split or paired reads discordant. However, there are a number of issues that arise with these structural variant callers. Firstly, they show very high rates of false discovery. That is, between 20 and 60% of the calls made have been shown to be false discoveries (English et al, 2015). Further, the calls are often imprecise in the breakpoints called (that is, they often indicate the breakpoint to be within a particular range). Finally, if there are structural variants that are inherited within a family, the calls are often not reported the same across individuals, both in terms of the location of the breakpoints and the size of the variants. This often requires additional work to be able to identify whether variants in similar locations are the same variants.

The high false discovery rate in calling structural variants, combined with the imprecision of calls, necessitates a tremendous amount of specialized bioinformatics work to narrow down the large number of candidate variants to be able to identify what could be the de novo structural variant that are associated with the disorder.

The BioGraph File Format



Here, we present results of detecting de novo variants using the BioGraph™ File Format.

Data in the BioGraph™ file format are stored in a way that allows rapid queries of the read data. Rather than spending considerable compute time to align against a given reference genome, and then analyze the data with different variant callers or assemblers, the data is processed up front to create BioGraph files, allowing any further queries to be both fast and flexible. This removes the need to use only fixed bioinformatics tools. Rather, it is possible to create queries to answer particular questions that return results quickly.

BioGraph™ files are best described as having all possible assembly information paths given the reads in a sample. In comparison, a reference holds information that is essentially flat (sometimes with additional contigs to allow for branching). The BioGraph file allows for rapid movement through the possible assembly paths.

The underlying format is based on a Burrows Wheeler Transform (http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.html). We extend the BWT, accommodating four different alternatives (nucleotides) at each location, enabling the data to be used as a graph based structure. Additionally, the FM index (Ferrangina and Manzini, 2000), further compresses the data and allows for rapid searching of the index. The rapid search feature has two functions:

- **Constant time traversal of the read graph for a sequence of any size**
- **The search for a subsequence is linear with the length of that sequence**

Accurate Structural Variant Calling Using the BioGraph Analysis Format

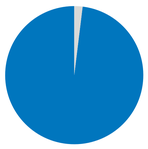
We developed a structural variant caller to run on top of BioGraph™. This method performs whole read overlap assembly on corrected, unmapped reads to detect SNVs, indels, and structural variants. Sequencing errors are corrected by base substitution within reads that contain k-mers that occur fewer than 4 times such that reads align to the de Buijn graph of all k-mers that occur at least 4 times (the graph representing true read sequences). Of these corrected reads, those that do not match the reference exactly are assembled into a discontiguous read overlap graph to capture sequence variation from the reference. Variants are mapped to human reference coordinates (GHCr37.p7) by walking the read overlap graph in both directions until an "anchor" read, where a continuous 70 bp matches the reference, denotes the beginning and end of each variant. Where a variant has more than one anchor, pairing information is used to determine the correct location of the anchor. The analysis presented here only includes variants classified as a deletion or an insertion.

A study by English et al. (2015) showed that, compared to a number of other variant callers, the BioGraph™ variant caller (previously called Anchored Assembly) showed sensitivity to structural variants with a low false discovery rate (< 3%).

Program	FDR	Sensitivity
CNVnator	80%	23%
BreakDancer	59%	42%
Delly	55%	31%
Crest	15%	35%
Pindel	32%	57%
SV-STAT	2%	16%
Tiresias	69%	8%
BioGraph™ structural variant caller	3%	34%

Table from English et al, 2015. Note that a more recent version of BioGraph™ variant caller showed a sensitivity of approximately 42%.

To show consistency of calls using this Format, we confirm structural variants identified by the caller Pindel using overlap assembly in an Ashkenazi Jewish Trio from the Personal Genome Project. Of 1,195 calls from Pindel that showed evidence in the reads for at least one individual (average size 252bp), all of these calls, except for 25 (2.1%), were consistent with mendelian inheritance in the BioGraph™.



98% (1,170 of 1,195) of variants discovered by Pindel that were observed to have read evidence in the BioGraph™ file format followed Mendelian inheritance.

In all cases that the variant was called in multiple samples, the variant was reported exactly the same.

To answer the question as to why some 2% of variants did not follow mendelian inheritance, a further analysis of coverage was completed using samples from a single individual that was sequenced at three different locations. Of approximately 298M 30-mers, 97.2% were the same across two samples. This indicates that approximately 2.8% of k-mers are seen in only one individual, suggesting a drop in coverage. This would explain why some variants do not follow mendelian consistency.

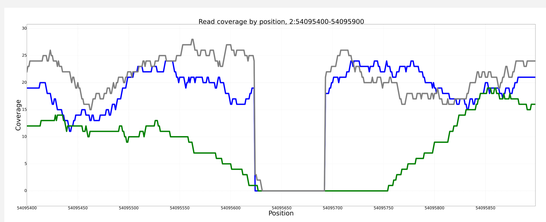


Figure 1. Graph of coverage for an individual sequenced at three centers. The location contains a deletion. There is a drop in coverage for one sample such that it may not be called as a deletion.

These analyses indicate that the BioGraph™ structural variant caller has a very low false discovery rate. Further, it indicates that variants are called the same across individuals, given sufficient coverage. This effectively addresses the issues of traditional bioinformatics analyses for trio analysis.

Observing Candidate De Novo Variants in a Real Case

We wanted to apply this method to a real case. The data in the following example comes from the Manton Center for Orphan Disease Research at the Boston Childrens Hospital.

The proband and both parents genomes were sequenced at 30x coverage using an Illumina HiSeq. Using the BioGraph™ structural variant caller (English et al, 2015), we discovered the structural variants present in this sample. We called 2,383 genetic changes that were either an insertion or deletion of greater than 50 base pairs in the proband. We then converted the data of the parents into the BioGraph format and queried them for evidence of these variants (and reference at the same location). Of the 2,383 structural variants, 98 showed, prima facie, evidence of being de novo. 18 of these variants were heterozygous in the proband. Although there was no evidence for the variant in either parent, there was a drop in coverage for at least one parent. Of the remaining 80 variants that were homozygous in the proband, the variant was present in at least one parent with the other parent either having no coverage at that location (68 variants) or low reference coverage (12 variants, < 9 reads). Overall, this is suggestive that these variants are more likely to be due to a lack of coverage than true de novo variants.

CONCLUSIONS

This analysis was completed in 18 hours. Of that, some 14 hours were used to compute the BioGraph™ format using the read data (from FASTQ files). In this case, it was not possible to confirm a true de novo structural variant. However, given that the method allows for accurate calling of structural variants, it is possible to rapidly detect these variants and, from the short list of candidates, rule out those that may be due to a lack of coverage. Applying this method to cases, in addition to existing methods, may result in solving more cases without dramatically increasing the use of resources or the time to return results.

FABRIC GENOMICS™

FORMERLY Omicia

Learn more at

www.fabricgenomics.com ■ info@fabricgenomics.com ■ 510.595.0800

Boston Children's Hospital — Until every child is well™

UTHealth — The University of Texas Health Science Center at Houston

Baylor College of Medicine

HGSC — HUMAN GENOME SEQUENCING CENTER