Poster # 2010

De Novo Assembly and Structural Variation Detection of Human Genomes using Single-Molecule Next-Generation Mapping and SV Call Validation by Inheritance and Orthogonal Measurements

A Hastie, T Anantharaman, T Liang, K Pham, M Saghbini, Ž Džakula, H Cao

[1]BioNano Genomics, San Diego, CA, USA

Abstract

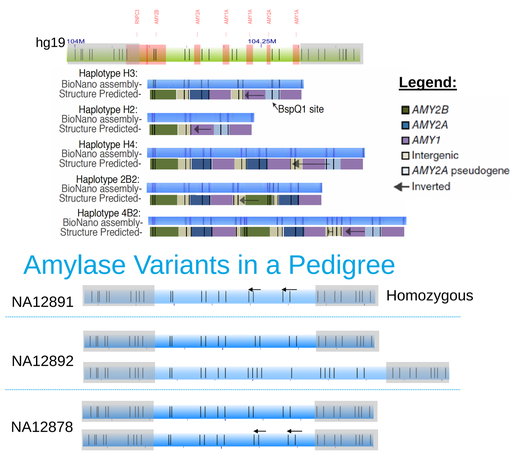
Structural variation analysis of human genomes typically relies on comparative analysis with a known reference. Because of this, the quality of the analysis is biased by the quality of that reference. To date, all human references contain unresolved structural variations within an incomplete assembly. In order to detect structural variations comprehensively, a method to assemble the genome without relying on a reference (a *de novo* assembly), is needed. Using the BioNano Genomics Irys® System, a single molecule genomic analysis system, we produced high-resolution genome maps that were *de novo* assembled while preserving the long-range structural information necessary for structural variation detection and analysis.

Here, the Genome in a Bottle (GIAB) reference trio of Ashkenazi Jewish decent (NA24385, NA24149, NA24143) has been *de novo* assembled using the Irys System. Structural variation (SV) analysis revealed insertions, inversions, and deletions, including large deletions in the *UGT2B17* gene (involved in graft versus host disease, osteopathic health, and testosterone and estradiol levels) in the mother's and son's genomes. We compared structural variants found in the son's genome (NA24385) by physically mapping them to those found in his parents' genomes. We found that deletion and insertion calls, as large as one kilobase pair and larger, found in the son's genome are also found in the parents' at a rate of 82% (deletions) and 91% (insertions).

We also used structural variation calls made with PacBio's long-read sequence assemblies to validate BioNano's structural variation calls, resulting in high confirmation rates in the 5 to 50 kilobase pair range. Thus, structural variation calls made using Irys System's genome mapping is an accurate method that can used to validate assemblies and SV's with other sequencing technologies

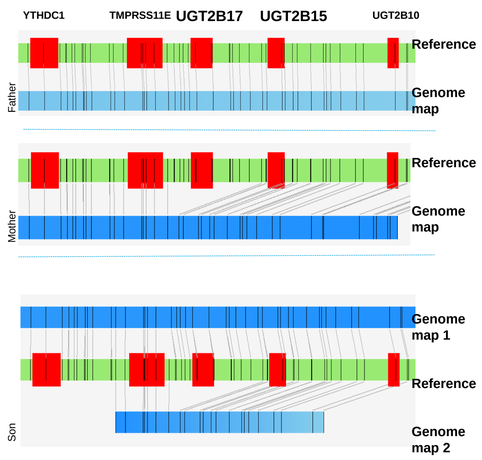
Methods

Amylase Structural Variants



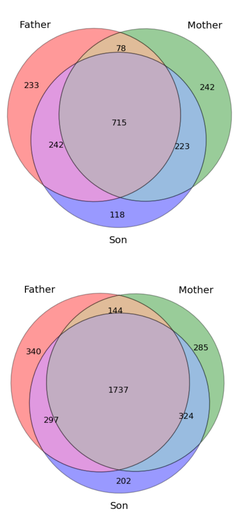
The amylase gene locus is polymorphic for structural variation. A previous study determined that copy number is associated with body mass index (BMI) but a more recent study questioned that conclusion. However, neither study investigated balanced structural variation in the population. At BioNano, through the analysis of numerous human genomes, we found that inversions occur frequently in the amylase locus, identifying new types of variants that may explain phenotypic observations. A formal study is needed to correlate our findings with biological outcome. The top graphic shows some of the copy number variants studied, while the bottom graphic shows a 2nd generation pedigree. In both cases, there are two alleles with the same copy number pattern but with different structures.

Gene Deletion in Mother and Son Genomes from the Ashkenazi Trio



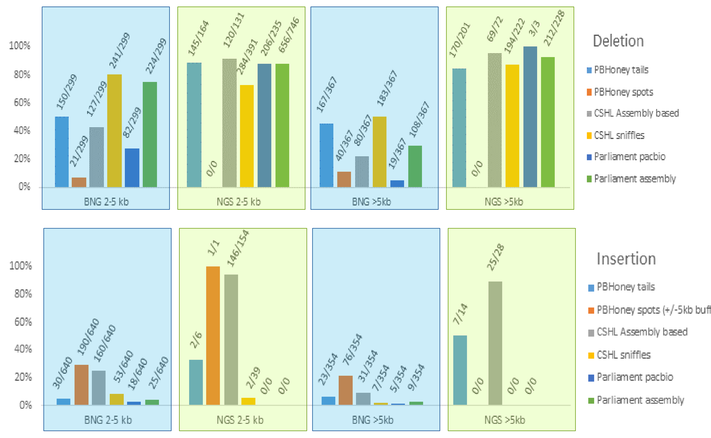
A 117 kbp deletion shows the missing UDP glucuronosyltransferase 2 family, polypeptide B17 gene (*UGT2B17*). Deletion of *UGT2B17* has been reported to result in better quality of osteopathic health as well as higher testosterone and estradiol levels. *UGT2B17* is believed to produce an important antigen involved in graft versus host disease (McCarroll).

Parents' Genomes' SV Calls Overlap in Son's Genome by an Average of 91%



Heritability of SV calls in a Trio. Deletion and insertion calls of 1 kbp and larger in the son's genome are found in the parents' at a rate of 90% and 92%, respectively, resulting in 2784 SV calls with expected inheritance patterns in the son's genome (NA24385).

Orthogonal Cross-Validation of NGM and NGM SV Calls



Comparison of SV calls using an orthogonal technology, PacBio long-read sequencing. We show that while BioNano smaller deletion calls also be called by various mapping-based algorithms (PBHoney, Sniffles, Parliament, etc.), larger deletions and insertions >2 kbp are however not efficiently detected by these algorithms.

Conclusions

The Irys System's next-generation mapping (NGM) technology is a unique, complementary and powerful method for detecting structural variations in the human genome. With NGM's ability to provide copy number information within genes, previous findings and assumptions should be reassessed so that new findings can be substantiated. Using NGM, large deletions, such as the missing *UGT2B17* gene in the Ashkenazi Trio's genomes, can be identified, which shows heritability of dynamics of genetic mutations. With NGM's ability to add to and confirm the human reference genome's structural architecture, it has proven itself to be a critical step in the completion of the gold standard human reference genome.

Reference

- Cao, H., et al., Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience (2014); 3(1):34
- Hastie, A.R., et al. Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex Aegilops tauschii Genome. PLoS ONE (2013); 8(2): e55864.
- Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
- Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.
- Usher et al., Structural forms of the human amylase locus and their relationships to SNPs, haplotypes, and obesity. Nature Genetics (2015); 47(8):921-5
- English. A, et al. Assessing structural variation in a personal genome—towards a human reference diploid genome. BioMed Central Genomics (2015); 16: 286
- Genome in a Bottle (2015). Next-generation Sequencing Structural Variation call data. Unpublished raw data.
- English, A. C., Salerno, W. J., & Reid, J. G. (2014). PBHoney: Identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics, 15(1), 180.
- Sedlazeck, Fritz, Cold Spring Harbor Laboratory (2015). Sniffles SVs and assembly based SVs. Unpublished raw data.
- DNA Nexus (2015). Parliament SVs. Unpublished raw data.