De Novo Assembly of the Genome-in-a-Bottle Reference Ashkenazi Trio, Structural Variation Discovery and Comparison with Other Individuals by Next-Generation Mapping

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

[1]BioNano Genomics, Research and Development, San Diego, CA

Abstract

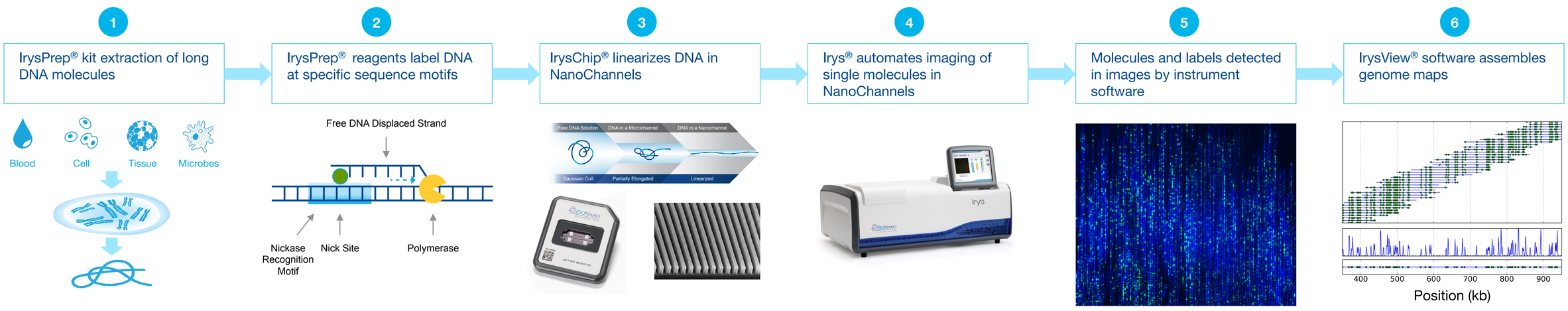
Structural variation analysis (SVA) of human genomes is usually a reference-based process and therefore biased and incomplete. In order to have a comprehensive analysis of structural variation, a *de novo* approach is needed. As a result of the remaining limitations of DNA sequencing and analysis technologies, it is not feasible to create high quality *de novo* assemblies of individuals for detecting and interpreting the many types of structural variation that are refractory to high-throughput or short-read technologies. Using a single-molecule genome analysis system, the Irys® System, we produced high resolution genome maps that were assembled *de novo*. These maps preserve long-range structural information necessary for structural variation detection.

The Genome in a Bottle (GIAB) reference trio of Ashkenazi Jewish decent (NA24385, NA24149, NA24143) has been *de novo* assembled by the Irys System. Structural variation analysis reveals 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 and son. We have also investigated the amylase locus in this trio as well as ~ 20 other individuals and have found at least 15 different structural variants. Human amylase genes have variable copy number and this variation is believed to have been evolved to adapt to increase starch intake. We were able to identify multiple copy neutral variants, such as inversions, in these individuals.

Background

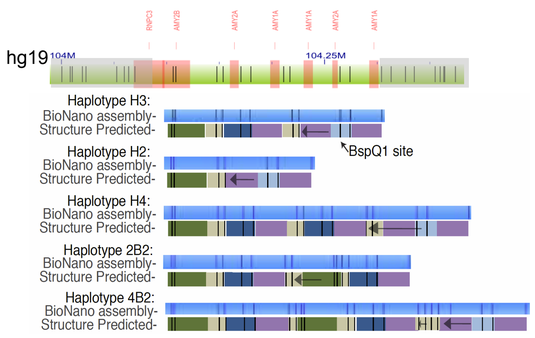
Structural variation in the human genome accounts for more bases changed in an average individual than SNPs, however, technologies for discovery and characterization have mostly been limited to array-based CNV detection and WGS. Arrays are considered low cost but have low resolution and known limitations. WGS generally is limited by its read length for SV detection. Therefore, the relationship between structural variation to human health and disease has been very difficult to study. Irys is commercialized for whole genome mapping by *de novo* assembly using very long single molecule reads. Because of its very long read length and its *de novo* approach, it is ideal for detection, discovery and interrogation of balanced and imbalanced structural variations. Here we demonstrate discovery and interrogation of two presumed health related SV polymorphic loci as well as introduce Irys' ability to *de novo* assemble long map contigs of unreferenced genomic material allowing new studies of these regions.

Methods

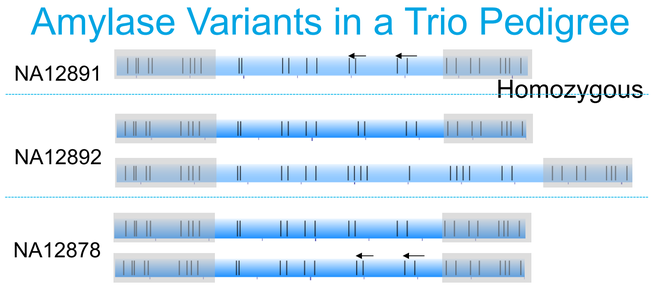


(1) Long molecules of DNA is labeled with IrysPrep® reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® NanoChannels and single molecules are imaged by Irys®. (4) Single molecule data are collected and detected automatically. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable and useful in assembly into genome maps. (6) Maps may be used in a variety of downstream analysis using IrysView® software.

Amylase Structural Variants

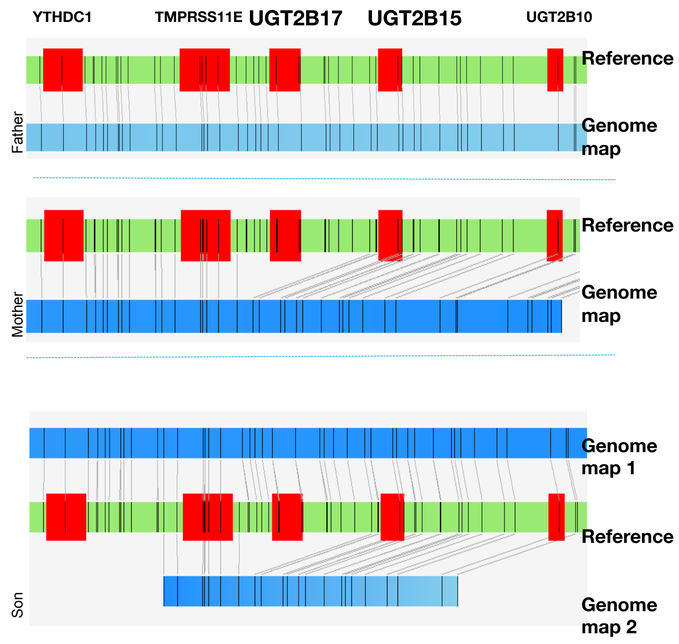


Amylase Variants in a Trio Pedigree



The amylase gene locus is polymorphic for structural variation. Previously, it has been determined that copy number has a relationship with body mass index (BMI) but recently that conclusion was questioned. Neither study was able to investigate the balanced structural variation in the population. We have analyzed many human genomes and find that inversions occur frequently in the amylase locus but no study has been conducted to try to correlate it with biologic outcome. The top panel shows various copy number variants studied, the bottom panel shows a trio pedigree. In this case, there are two alleles with the same copy but with different structures.

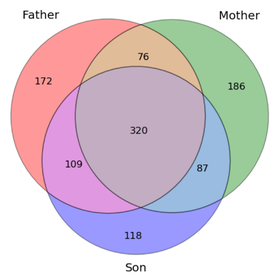
Gene Deletion in Mother and Son from Ashkenazi Trio



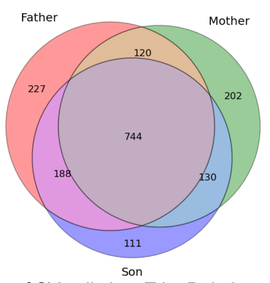
A 117 kb deletion removes UDP glucuronosyltransferase 2 family, polypeptide B17 (UGT2B17). Deletion of UGT2B17 has been reported to result in increased osteopathic health as well as higher testosterone and estradiol levels. UGT2B17 is believed to produce an important antigen involved in graft versus host disease (McCarroll).

Parental Overlap of SV Calls in NA24385

Deletions



Insertions



Heritability of SV calls in a Trio. Deletion and insertion calls 1 kb and up in the son are found in the parents at a rate of 82% and 91% respectively resulting in 1578 high confidence SV calls in the son (NA24385).

Conclusions:

BioNano next generation mapping presents a fast and reliable solution for structural variation detection and discovery in human genomes, which provides copy number and balanced structural information in the amylase gene region. Using BioNano's next generation mapping *de novo* assembly, large deletions that include UGT2B17 are identified. The *de novo* assembly techniques are needed for showing the whole genome, where one percent of genome map DNA is missing from the reference. For more information about next-generation mapping, also see Posters #1832T, #3118T, #2721W and #1632F.

Reference:

1) Cao, H., *et* al., Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience (2014); 3(1):34
2) 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.
3) Lam, E.T., *et* al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
4) Xiao, M., *et.* al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.
5) 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