De Novo Assembly and Structural Variation Discovery in in Human Disease and Non-Disease State Genomes Using Extremely Long Single-Molecule Imaging

Authors: Alex Hastie[1], Ernest Lam[1], Tiffany Liang[1], Andy Pang[1], Saki Chan[1], Han Cao[1]

[1]BioNano Genomics, San Diego, CA, 92121, USA

Abstract

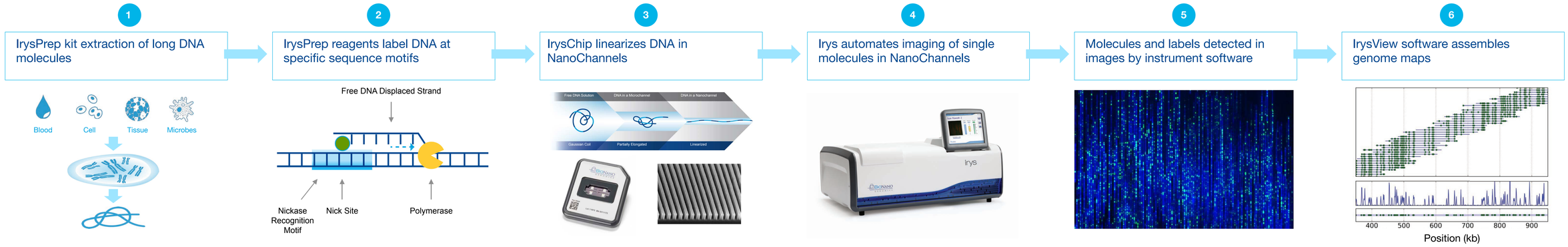
Structural variation analysis (SVA) of human genomes is usually a reference based process and therefor biased and incomplete. In order to have a comprehensive analysis of structural variation, a *de novo* approach is needed. *De novo* genome assemblies using only short read data are generally incomplete and highly fragmented due to the intractable complexity found in the human genome. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses. As a result of the remaining limitations of DNA sequencing and analysis technologies, it is not feasible to create high quality assemblies of individuals to detect and interpret the many types of structural variation that are refractory to high throughput or short-read technologies.

We present a single molecule genome analysis system (Irys®) based on NanoChannel Array technology that linearizes extremely long DNA molecules for direct observation. This high-throughput platform automates the imaging of single molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness for unambiguous assembly of complex genomes. High resolution genome maps assembled *de novo* preserve long-range structural information necessary for structural variation detection and assembly applications. Dozens of human genomes have been *de novo* assembled by Irys to date, including cancer genomes. Structural variation analysis reveals insertions, deletions, inversions and translocations. We have generated genome maps for two trios, CEPH trio (NA12878, NA12891 and NA12892) and an Ashkenazi jewish trio (AJ; NA24385, NA24143 and NA24149). From these genome maps, we detect hundreds of structural variants, including large deletions that delete genes in the mother and son from the AJ trio. We have also investigated the amylase locus in both trios 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 are able to identify multiple copy neutral variants, i.e. inversions, for each for the same copy number variants. Each genome shows additional dramatic structural variation, including many megabases of variation within genomic regions not included in the public reference genome assembly, underscoring the need for more *de novo* approaches to genome analysis.

Background

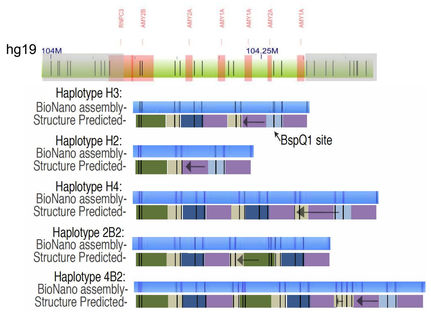
Structural variation in the human genome accounts for more bases changed in an average individual than SNPs, however, technologies for discovery and characterization have mostly been limited to array based CNV detection, WGS. Arrays are inexpensive but low resolution and error prone. WGS generally is limited by it's read length for SV detection. Therefore, the relationship between structural variation to human health and disease has been very difficult to study. Irys has been commercialized for whole genome mapping by *de novo* assembly using very long single molecule reads. Because of it's very long read length and it's *de novo* approach, it is ideal for detection, discovery and interrogation of balanced and imbalanced structural variation. Here we demonstrate discovery and interrogation of two presumed health related SV polymophic 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

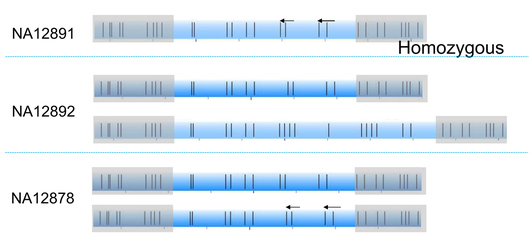


(1) Long molecules of DNA are labeled with IrysPrep® reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® nanochannels and single molecules are imaged by Irys. (4) Single molecule 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

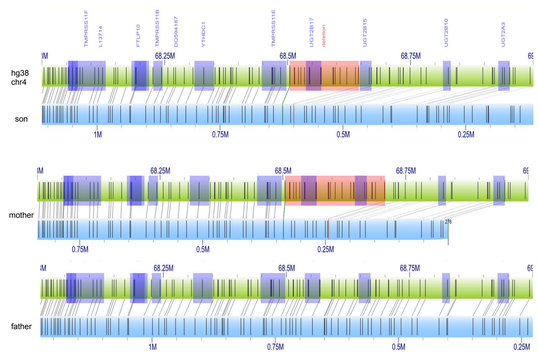


Amylase Variants in a Trio Pedigree



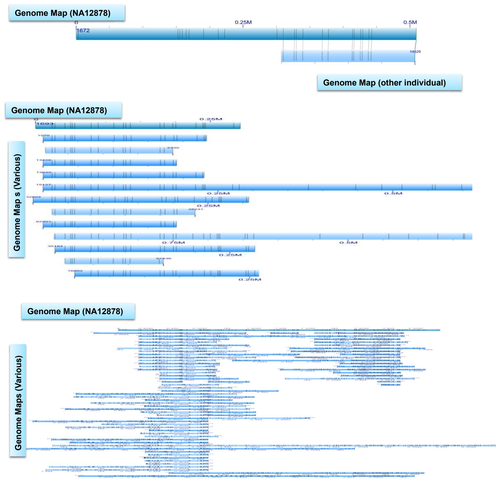
The amylase gene locus is polymorphic for structural variation. Previously, it has been determined that copy number has a relationship with BMI but recently that conclusion was questioned. Neither study was able to investigate the balanced structural variation in the population. We have analyzed many human genomes and find that inversions occur frequently in the amylase locus but no study has been conducted to try to correlated 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).

De novo Assembly Produces >30Mb of DNA Absent from References



In the NA12878 genome map, 26Mb from 51 maps do not align to the GRCh38 reference. When comparing the 51 unmapped maps to other genome maps, most align to genome maps assembled from multiple different individuals, examples are shown on the left. Unplaced NGS sequence can be anchored to the unmapped genome maps.

NA12878 Map ID	Length	Occurrence in other genome maps from other individuals
1672	509,474	1
9536	197,386	1
1522	274,787	2
7550	169,684	2
1576	280,393	3
2592	100,713	3
4723	123,662	3
1352	286,682	4
1793	136,730	4
1011	697,855	5
2752	121,323	5
7553	129,557	5
1702	366,784	6
14469	479,914	6
1019	593,042	8
1471	706,429	8
5274	143,866	8
1071	449,186	9
1219	310,568	9
1330	152,780	10
1409	304,891	10
1180	450,435	11
1601	277,510	11
5489	120,344	11
1172	440,745	12
1292	715,096	12
1693	292,896	12
1406	292,504	13
1459	639,244	13
1475	425,298	14
1279	687,797	15
1336	576,921	15
1017	529,133	16
1080	857,117	18
1569	613,189	26
1103	963,954	29
9698	548,212	29
972	607,283	31
572	1,118,409	43
1061	1,264,805	44
1245	529,630	44
920	2,493,496	61
713	2,979,866	98
99	946,866	214

Conclusions:

- Genome mapping is a fast and reliable method for SV detection and discovery in human genomes.
- Genome mapping provides copy number and balanced structural information in the amylase gene region that.
- Genome mapping identified large deletions that includes UGT2B17.
- One percent of genome map DNA is missing from the reference, de novo assembly approaches are need to study the whole genome.

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) Falchi et al., Low copy number of the salivary amylase gene predisposes to obesity. Nature Genetics (2014); 46(5):492-498
6) Usher et al., Structural forms of the human amylase locus and their relationships to SNPs, haplotypes, and obesity. Nature Genetics (in press)
McCarroll et al