Abstract

Comprehensive genome structural characterization requires whole genome de novo assembly of DNA fragments that span repeat regions in the genome. Next-generation mapping (NGM) using BioNano Genomics’ Irys® System is based on the direct visualization of extremely long genomic fragments and it offers a high-throughput and genome-wide method, to interrogate genome structural variations (SVs) in the range of sub kilobase to megabase pairs, thus providing the most complete and structurally intact assembly available.

Here we characterize the quality of NGM assemblies for the use of detecting structural variations in human genomes. Because there is no ground-truth human genome sequences available, all assemblies have errors, we have produced simulated molecules that faithfully reproduce characteristics of BioNano’s single molecule datasets. These simulated molecules can then be de novo assembled and aligned to a reference human genome to find structural differences. We have introduced structural variations into our simulated molecules in order to measure the sensitivity and accuracy of the method across different variant types and sizes.

We report on insertions and deletions between 500 bp and 500 kbp. This analysis shows that sensitivity ranges from 77% and 69% at 1.5 kbp up to 94% and 93% at 5 kbp for insertions and deletions, respectively. Sensitivity remains above 80% up to 150 kbp for insertions and begins to drop as the SV size approaches molecule (read) lengths. Importantly, the insertions are still detected but not correctly classified due to contiguity breaks within the inserted sequence. We also report on translocation sensitivity and accuracy. Here we evaluated translocation breakpoint detection in a heterozygous scenario and show that sensitivity is 76% for all tested translocation breakpoints with most of the remaining translocations being truncated and classified as “end” SVs. For all SV types we show biological validation that agrees well with the simulated data.

This method is being adopted for cytogenetically characterizing genomes of patients with inherited and spontaneous genetic disease such as developmental disorders and cancer. The high speed, sensitivity and accuracy when compared to sequencing based methods and orders of magnitude higher resolution when compared to traditional cytogenetics make Irys an attractive choice.

Methods

1. IrysPrep kit extraction of long DNA molecules
2. IrysPrep reagents label DNA at specific sequence motifs
3. IrysChip linearizes DNA in NanoChannel arrays
4. Irys automates imaging of single molecules in NanoChannel arrays
5. Molecules and labels detected in images by instrument software
6. IrysView® software assembles optical maps

(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® using NanoChannel arrays 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 analyses using IrysView® software.

Insertion and Deletion Sensitivity and Positive Predictive Value – Simulated Molecules

<table>
<thead>
<tr>
<th>SV Size Range</th>
<th>Number of Simulated Translocations</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 bp</td>
<td>100</td>
<td>76%</td>
</tr>
<tr>
<td>1 kbp</td>
<td>100</td>
<td>93%</td>
</tr>
<tr>
<td>&gt;1 kbp</td>
<td>100</td>
<td>79%</td>
</tr>
</tbody>
</table>

Genomes were assembled from simulated molecules containing heterozygous insertions and deletions. Overall sensitivity was 76% but including only Translocations that are at least 200 bp, as is usually the case in cancer genomes, sensitivity was up to 82%. Translocation detection performed better outside of segmental duplication but still had 60% sensitivity even within segmental duplications. Sex chromosomes contained more chimeric genome maps when compared to autosomes, presumably as a result of pseudo-autosomal regions.

Translocation Sensitivity and Positive Predictive Value – Simulated Molecules

<table>
<thead>
<tr>
<th>SV Size Range</th>
<th>Number of Simulated Translocations</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 bp</td>
<td>100</td>
<td>76%</td>
</tr>
<tr>
<td>1 kbp</td>
<td>100</td>
<td>93%</td>
</tr>
<tr>
<td>&gt;1 kbp</td>
<td>100</td>
<td>79%</td>
</tr>
</tbody>
</table>

Genomes were assembled from simulated molecules containing heterozygous translocations and transpositions. Overall sensitivity was 76% but including only Translocations that are at least 200 bp, as is usually the case in cancer genomes, sensitivity was up to 82%. Translocation detection performed better outside of segmental duplication but still had 60% sensitivity even within segmental duplications. Sex chromosomes contained more chimeric genome maps when compared to autosomes, presumably as a result of pseudo-autosomal regions.

Biological Validation: Translocations

In a blind study, we analyzed three leukemia samples where translocations were found using fluorescence in situ hybridization. By using NGM we were able to correctly identify all three translocations and map the breakpoints to ~5 kbp regions.

Conclusions

De novo detection of insertions, inversion and translocations is inefficient and inconsistent by commercially available technologies. Next-generation mapping (NGM) using the Irys System, is efficient, cost effective, sensitive and specific for structural variation analysis. Sensitivities for Insertions and Deletions are within 80-95% across the entire genome above 1.5 kbp in simulation experiments. We further validated our insertion and deletion calls in a biological sample and those in the range between 1-20 kbp were validated by Pacific Biosciences sequencing. Translocation detection is very specific and sensitive in both in silico and biological validation experiments. Because of its speed and accuracy, NGM is being adopted by researchers for the characterization of clinical genomes.

References


*We would like to thank Jonathan Diver of Genoptix, Carlsbad, CA, collaborative efforts in experimental design and interpretation.