Genoox

# A Novel Approach for Structural Variant Calling: Combining Data from Whole Genome Next-generation Sequencing and Optical Mapping

Yuval Porat, Oron Lev, Moshe Einhorn, Odem Shani, Eric Vilain, Hayk Barseghyan, Surajeet Bhattacharya, Nurit Paz-Yaacov
Genoox, Tel Aviv, Israel

## ACCEPTED ABSTRACT

Structural variant (SV) detection via short-read sequencing is typically characterized by high false-positive rates and considerable uncertainty regarding variant breakpoint position, owing to the size of alterations typically far exceeding the read length. Thus, alignment patterns between short reads and the reference genome often can provide only limited evidence for the existence of a variation. In addition, variant calling and genotyping employ probability inference surrounding combined alignment signals, which carries inherent uncertainty surrounding exact breakpoint positions and confidence. Optical mapping technology for SV detection results in high sensitivity and specificity but often carries limited certainty of breakpoint positions.

We validated a novel joint pipeline for SV detection that integrates next-generation sequencing (NGS) raw data with optical mapping-based SV calls using an advanced reference graph structure. By applying the reference graph structure throughout the analytical pipeline, evidences from both technologies can be considered simultaneously to provide increased breakpoint precision and confidence. The novel SV-calling pipeline was validated for deletions and insertions larger than 500 base pairs (bps), using sample data and high-confidence SV truth sets available for NA12878 in Genome in a Bottle (GIAB; http://jimb.stanford.edu/giab), by comparing NGS-only, optical-mapping-only, and combined technologies calling.

Per optical-mapping-only data, 817 deletions and 1,670 insertions had high breakpoint uncertainty (tens of thousands bp). With application of the combined technologies, the uncertainty of accurate breakpoint positions for the 0-75th percentile of the 756 (92.5%) deletions and 1,248 (75%) insertions called was reduced to only 4bp and 140bp, respectively. Further, results of a specificity evaluation using in silico simulations on regions that did not overlap the GIAB truth set indicated 99.3% (971/978) specificity for deletions and 94.3% (910/965) for insertions. As such, the respective false-positive rates were 0.7% (7/978) and 5.7% (55/965).

A novel joint pipeline integrating NGS raw data with optical mapping provided SV calls characterized by precise breakpoint positions and low false-positive rates, suggesting the advanced graph structure could have considerable utility in clinical practice.

***NOTE: The data shown in this poster reflect further refinements to the joint pipeline since the time of abstract submission. No conclusions derived from this validation experiment have changed.***

American Society of Human Genetics
2018 Annual Meeting
October 16–20, 2018 ■ San Diego, CA

Genoox

# A Novel Approach for Structural Variant Calling: Combining Data from Whole Genome Next-generation Sequencing and Optical Mapping

Yuval Porat, Oron Lev, Moshe Einhorn, Odem Shani, Eric Vilain, Hayk Barseghyan, Surajeet Bhattacharya, Nurit Paz-Yaacov
Genoox, Tel Aviv, Israel

*NOTE: The data shown in this poster reflect further refinements to the joint pipeline since the time of abstract submission. No conclusions derived from this validation experiment have changed.*

## Introduction

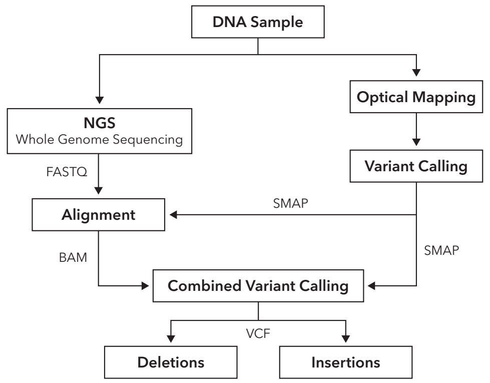
- Genomic structural variants (SVs) are recognized as major sources of genomic diversity[1-3]
- SVs such as copy number variations (CNVs) are responsible for a rapidly increasing number of genomic disorders,[3,4] including –
  - Mendelian diseases
  - Many common complex traits including autism and schizophrenia
- Current methods for detecting SVs have limitations
  - Short-read sequencing via next-generation sequencing (NGS) – typically characterized by high false-positive rates and considerable uncertainty regarding variant breakpoint position[5]
  - Optical mapping (OM) – although relatively sensitive and specific, certainty of breakpoint positions typically limited[6]
- Bionano OM relies on technology whereby DNA is labeled, linearized in a specialized nano-channel, and imaged on a single-molecule level; data from multiple DNA molecules are analyzed to map the genome structure and call SVs
- We developed and validated a joint pipeline for SV detection by integrating NGS raw data with OM-based SV calls

## Methods

### Joint Pipeline for SV Calling

- The joint pipeline for SV detection integrates NGS raw data with OM-based SV calls using an advanced reference graph structure **(Figure 1)**

**Figure 1. Joint pipeline (NGS + OM) for SV calling**



- Variants identified as insertions and deletions by Bionano Genomics' Saphyr™ System were employed
  - During alignment of whole-genome NGS reads, OM parameters are employed by a Genoox-developed graph aligner to create the reference graph structure
  - Subsequently, during variant calling, OM parameters are used as signals to validate and increase the confidence of potential variants
  - The final product is a variant call format (VCF) containing deletions and insertions, where each called variant is matched with a Bionano variant
  - For each called variant, the uncertainty interval is calculated for each breakpoint (1 breakpoint for insertions; 2 breakpoints for deletions); the true variant breakpoint falls within the uncertainty interval

### Joint Pipeline Validation

- The joint pipeline for SV detection was validated using Bionano OM parameters and NGS data for sample NA12878 from the 1000 Genomes Project Phase 3 (1K Genomes; internationalgenome.org)
  - Only Bionano variants >500 base pairs (bps) and with Bionano confidence score ≥0.5 employed
- Resulting OM variant list included 1,115 unique deletions and 2,243 unique insertions
- Deletions and insertions – compared with the 1K Genomes variant calls
  - >500 bps
  - ≥70% reciprocal overlap with the variants in the Bionano OM and Genoox joint pipeline sets
- Specificity was determined using a simulated OM dataset of false-positive variants

## Results

### Deletions

- The Genoox joint pipeline called **977 (87.6%)** of the 1,115 Bionano OM deletion variants
  - Deletion sizes called by the joint pipeline were generally consistent with the sizes called by Bionano's OM caller **($R^2$ = 0.98)**
  - Most of the variation in sizes was observed in the shorter variants **(Figure 2)**
- Uncertainty of Bionano OM **1,115** deletions **(Figure 3)**
  - Calculated by combining the uncertainty interval of both breakpoints
  - 90% of the variants' uncertainty is between 2 × $10^3$ and 3 × $10^4$ bps long.
  - No correlation with the deletion size
- Uncertainty of Genoox pipeline 977 deletions **(Figure 3)**
  - Calculated by combining the uncertainty interval of both breakpoints
  - Uncertainty of **931 variants (83.5%** of Bionano OM-only variants) reduced by ≥2-fold
  - Uncertainty of **690 variants (61.9%)** reduced to <100 bps, providing a relatively very accurate location
  - Minimal uncertainty for a single breakpoint: 4 bps

**Figure 2. Correlation of deletion sizes called by Genoox joint pipeline and Bionano OM**

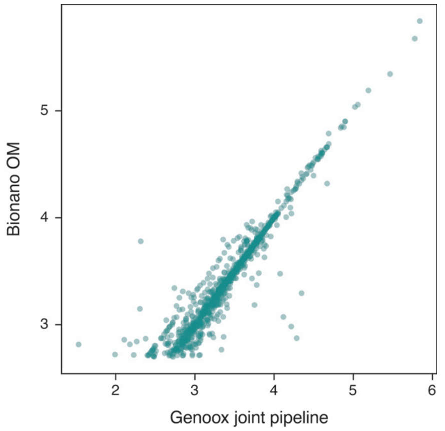
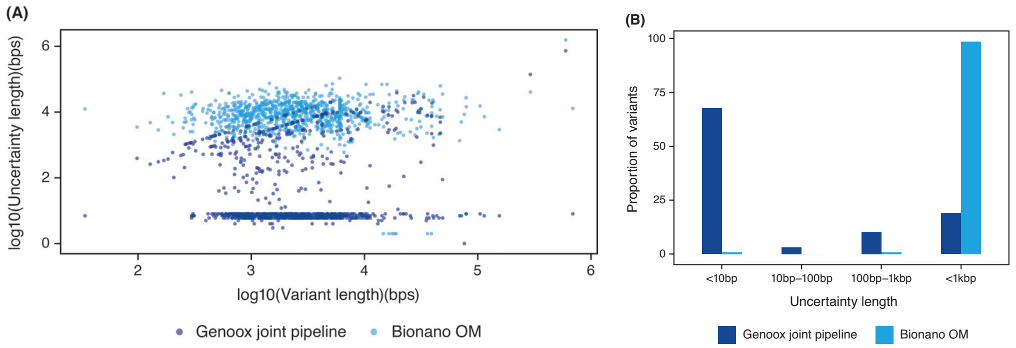


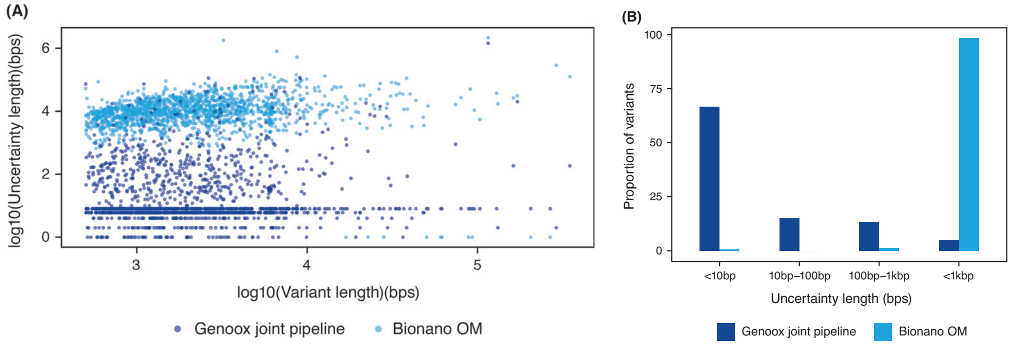
**Figure 3. Uncertainty of called deletions (N=977); relationships between (A) variant uncertainty and variant lengths and (B) proportions of variants and uncertainty size**



### Insertions

- The Genoox pipeline called **1,668 (74.4%)** of 2,243 unique high-confidence insertions called by Bionano OM
- Uncertainty length of Bionano OM insertions **(Figure 4)**
  - Determined by the size of the area, indicated by OM, in which the variant is located
- Uncertainty length of Genoox pipeline insertions **(Figure 4)**
  - For **1,637 (73.0%)** of 2,243 insertions, breakpoint uncertainty reduced by ≥2-fold
  - For **1,362 (81.7%)** of 1,668 insertions called, breakpoint uncertainty reduced to <100 bps

**Figure 4. Uncertainty of called insertions (N=1,668); relationships between (A) variant uncertainty and variant lengths and (B) proportion of variants and uncertainty size**



### Comparison to 1K Genomes Truth Set

- Relative to 1K Genomes variant calls (variants >500 bps, ≥70% reciprocal overlap between 1K Genomes variants and Bionano OM-only variants and between 1K Genomes variants and Genoox pipeline variants) **(Table 1)**:

**Table 1. Overlap of 1K Genomes variant calls for sample HG001 with Bionano OM-only and Genoox pipeline calls**

| | Deletions | Insertions |
|---|---|---|
| **1K Genomes truth set, N=** | 679 | 113 |
| **Bionano OM-only calls, n (%)** | 593 (87.3%) | 92 (81.4%) |
| **Genoox pipeline calls with improved precision, n** | 516 | 84 |
| % 1K Genomes truth set | 76.0% | 74.3% |
| % Bionano OM-only set | 87.0% | 91.3% |

### Pipeline Breakpoints Refinement Specificity

- The Genoox pipeline demonstrated robust breakpoint specificity: **98.2%** for deletions and **90.1%** for insertions **(Table 2)**

**Table 2. Breakpoint specificity of the Genoox pipeline**

| Variant type | Simulated Bionano OM false-positive signals[1] | Genoox pipeline variants called | Specificity[2] |
|---|---|---|---|
| **Deletions** | 1,013 | 18 | 98.2% |
| **Insertions** | 954 | 860 | 90.1% |

[1] Simulated by constructing regions that mimic a Bionano false positive call, by making sure they don't intersect any truth variant or any real Bionano variant

[2] Breakpoint specificity = 100 – [(# of false variants found / # of simulated variants) * 100]

## CONCLUSIONS

- By combining data generated from two variant calling techniques – NGS and OM – the Genoox pipeline for SV detection described herein has demonstrated robust sensitivity and specificity
- The breakpoint uncertainty interval of most deletions and insertions called was reduced from thousands of bps to the level of single bps, while also maintaining the breakpoint specificity and confidence of called variants
- Reducing the breakpoint uncertainty interval of these variants allows for:
  - More accurate prediction of the SV's exact effect on the surrounding genomic element
  - Identified SVs to be used in clinical analysis, interpretation, and decision making
  - Separation and classification of the hundreds of SVs present in the human genome
  - Design of variant confirmation by alternative methods (e.g., Sanger sequencing, real-time polymerase chain reaction), by narrowing the area that needs to be tested

**References**

1. Pendleton M, et al. *Nature Methods* 2015;12:780.
2. Huddleston J, et al. *Genome Res* 2017;27:677-85.
3. Stankiewicz PÇ, et al. *Annu Rev Med* 2010;61:437-55.
4. Lee H, et al. *JAMA* 2014;312:1880–7.
5. Guan P, et al. *Methods* 2016:102:36-49.
6. Li L, et al. *Genome Biol* 2017;18:230.