Sensitive detection of low-allele fraction structural variants in clinical cancer samples

Ernest T Lam, Yannick Delpu, Andy WC Pang, Thomas Anantharaman, Jian Wang, Javier Velazquez-Muriel, Tom Wang, Dong Zhang, Rayan Massoud, Scott Way, Alex R Hastie, Mark Borodkin
Bionano Genomics, San Diego, California, United States of America

Abstract

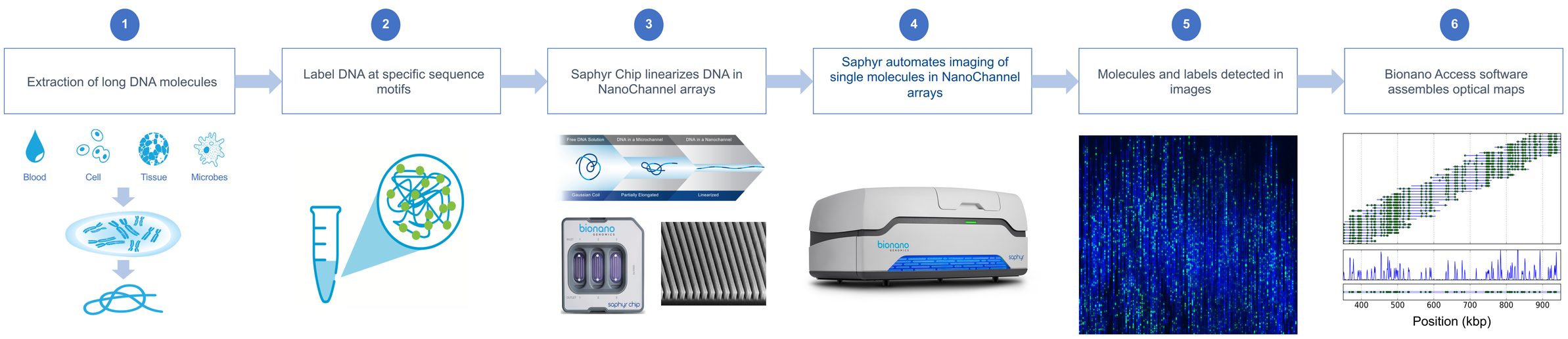
Tumors are often comprised of heterogeneous populations of cells, with certain cancer-driving mutations at low allele fractions in early stages of cancer development. Effective detection of such variants is critical for diagnosis and targeted treatment. However, typical short sequence reads are limited in their ability to span across repetitive regions of the genome and to facilitate structural variant (SV) analysis. Based on specific labeling and mapping of ultra-high molecular weight (UHMW) DNA, we developed a single-molecule platform that has the potential to detect disease-relevant SVs and give a high-resolution view of tumor heterogeneity.

We have developed a pipeline that effectively detects structural variants at low allele fractions. It includes single-molecule based SV calling and fractional copy number analysis. Preliminary analyses using simulated data and well-characterized cancer samples showed high sensitivity for variants of different types at as low as 5% allele fractions with reasonable genomic coverage easily collectable on a Bionano Saphyr Chip. The candidate variants are then annotated and further prioritized based on control data and publically available annotations. The data are imported into a graphical user interface tool that includes new visualization features (such as dynamic variant filtering, Circos diagrams, and report generation) for interactive visualization and curation. Together, these components allow for efficient analysis of any cancer genome of interest.

Background

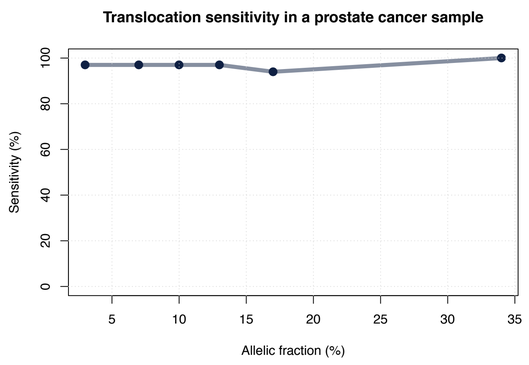
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. The Saphyr™ system provides direct visualization of long DNA molecules in their native state, bypassing the statistical inference needed to align paired-end reads with an uncertain insert size distribution. These long labeled molecules are *de novo* assembled into physical maps spanning the entire diploid genome. The resulting provides the ability to correctly position and orient sequence contigs into chromosome-scale scaffolds and detect a large range of homozygous and heterozygous structural variation with very high efficiency.

Methods



(1) Long molecules of DNA are labeled with Bionano reagents by (2) incorporation of fluorophores at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the Saphyr Chip using NanoChannel arrays (4) Single molecules are imaged by Saphyr and then digitized. (5) Molecules are uniquely identifiable by distinct distribution of sequence motif labels (6) and then assembled by pairwise alignment into *de novo* genome maps.

Rare Variant Pipeline



Data from a primary prostate cancer sample were mixed with control data at different fractions (1000X effective coverage).

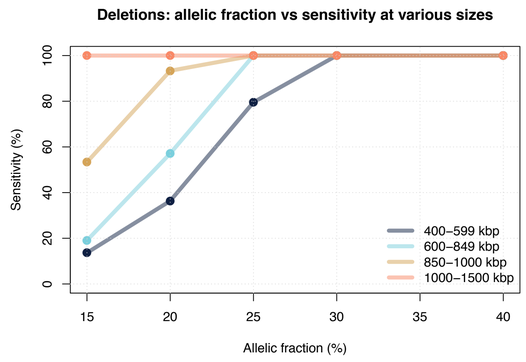
The Rare Variant Pipeline (RVP) provides ***high sensitivity*** for variants at ***low allele frequencies*** that are prevalent in heterogeneous samples such as cancers and samples with genetic mosaicism.

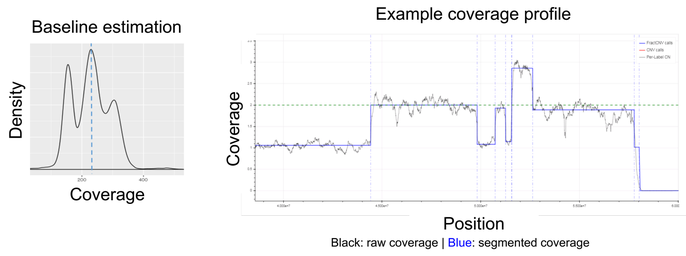
It performs a split read analysis:
1. Single molecule-based SV calling and clustering
2. Consensus generation for each cluster
3. Final confirmation of SVs using consensus maps

Based on analysis of simulated data, > 90% sensitivity at 300X effective coverage at 5% variant allele frequency:

- Insertions between 5 – 50 kbp
- Deletions > 5 kbp
- Translocations (or transpositions > 70 kbp)
- Inversions > 100 kbp
- Duplications > 150 kbp

Fractional copy number analysis



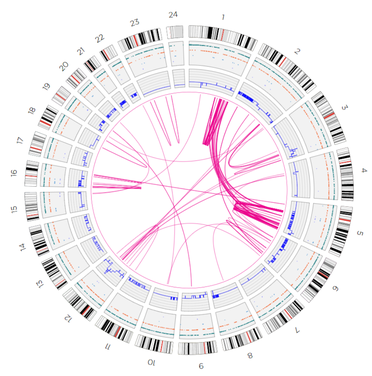


A new fractional copy number (CN) analysis pipeline was implemented to detect events in genomes with multiple CN state changes and events at lower allele frequencies.

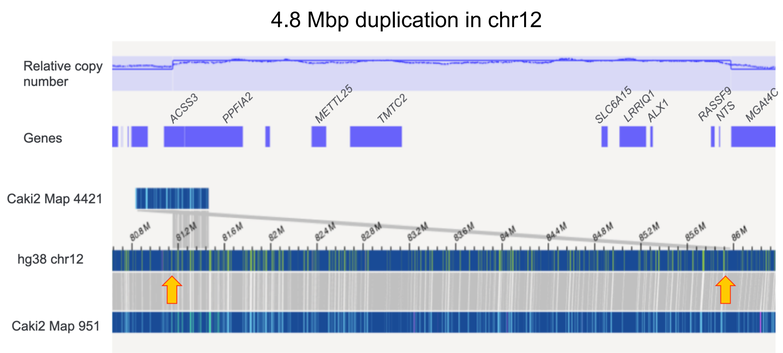
We implemented a new baseline estimation procedure to use the coverage mode as the baseline and a new statistical model to test deviations from the baseline (bottom left figure).

We incorporated a new copy number segmentation procedure (Wild Binary Segmentation) that is efficient and effective for handling multiple changes points (bottom right figure).

Data visualization



Track 1 (outer most): ideaogram
Track 2: insertion/deletion, inversion breakpoint and duplication calls
Track 3: copy number profile,
Inner circle: translocation breakpoint calls.



Circos visualization of SVs in Caki2 cell line *(top)* and a duplication detected in the sample *(bottom)*.

Advanced visualization options (such as Circos plots) are developed to facilitate exploratory and interactive data analysis. In this sample, we detected a large duplication that overlapped multiple genes and was not found in the Bionano control database.

Conclusions

Understanding of the genome structure is important for disease studies. However, typical short sequence read sequencing is expensive at coverage depths needed for detection of variants in rare clones. It is also limited in its ability to span across repeats in the genome and this results in high error rates in structural variant (SV) analysis. Bionano provides a streamlined workflow for efficient and comprehensive analysis of the genome structure.

We developed a suite of bioinformatics tools (collectively called Bionano Solve) that are integrated with Bionano Access, which serves as the graphical user interface for running various downstream analyses. Access's data visualization tools are interactive, and they allow for curation of SVs of interest. Using these tools, researchers can map their genomes of interest in a single assay for under $500 per sample.

Reference

Mak AC et al. Genome-wide structural variation detection by genome mapping on nanochannel arrays. Genetics (2016); 202: 351-62.
Cao H et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. GigaScience (2014); 3(1):34