Detection and Characterization of Copy Number Variations and Complex Genomic Rearrangements in Human Subjects Using Nanochannel Genome Mapping Technology

Ž. J. Džakula, W. Andrews, H. Dai, T. Anantharaman, A. Hastie, H. Cao
Computational Biology, BioNano Genomics, San Diego, California, USA

Abstract

Alterations of genomic DNA, ranging from simple copy-number variations to complex genomic rearrangements, often induce anomalous phenotypes. The ability to observe DNA molecules long enough to span the affected genomic regions is essential for fully realizing the diagnostic potential of these structural variants. Irys, a high-throughput genome analysis platform based on Nanochannel Array technology, is uniquely suited for such task as it linearizes extremely long (hundreds of kilobases in size) DNA molecules for direct imaging. The detected molecules are used to assemble high-resolution *de novo* consensus genome maps, enabling detection of long-range genomic structural rearrangements. In addition, copy number variations are analyzed by aligning the observed single molecules of genomic DNA to the reference and by normalizing the resulting coverage profile. By combining the structural variations found in human *de novo* assemblies with the features detected in the corresponding normalized copy-number profiles, we are able to comprehensively characterize whole chromosome aneuploidies, microdeletions, intra- and inter-chromosomal translocations, and inversions. The results are illustrated in cancer samples and in euploid subjects.

Long DNA molecules, specifically labeled with fluorescent tags and linearized on the Irys platform, can be recorded and mapped to the reference genome. The resulting raw depth profile consists of a sequence of counts of aligned molecules that cover each label along the reference. While conveying the copy number information with ~1 data point per 10kb, the raw depth profile also contains stochastic and systematic fluctuations, which may conceal the biological sources of variability. To increase the accuracy of copy number calls, it is essential to reduce the systematic component of the technical variability. Our approach is based on the observation that the success of molecule alignment to portions of reference that contain few labels increases with molecular length, while label-rich regions successfully attract even shorter molecules. The ability of any given reference label to attract molecules during alignment reflects both the descriptors specific to the reference (pattern of labels surrounding the label in question), and the characteristics of the sample preparation (distribution of molecular lengths). We separate the two components of bias by taking advantage of population data recorded on a set of euploid human samples. Multiple measurements of raw depth values, measured for the same label in different samples, are analyzed as a function of the average molecule length per sample using linear regression. The regression coefficients describe the expected counts per label in the absence of bias (intercept, *I*) and the susceptibility of the given label to biases induced by variations in the sample-specific molecular length distribution (slope, *S*). Thus trained, the model parameters are used to predict copy number values per label in new samples. Perturbation-based error propagation analysis is used to filter out less reliable labels.

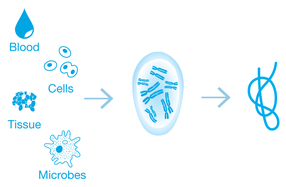
Background

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. Instead, Irys technology provides direct visualization of long DNA molecules in their native state, avoiding the statistical assumptions that are normally used to force sequence alignments of low uniqueness elements. The resulting order and orientation of sequence elements are demonstrated in anchoring NGS contigs and structural variation detection.

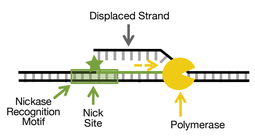
Methods

(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.

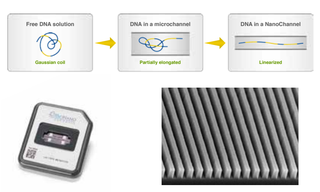
1) IrysPrep Kit extraction of long DNA molecules



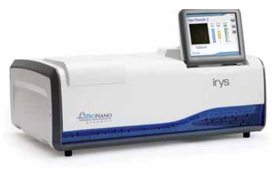
2) IrysPrep reagents label DNA at specific sequence motifs



3) IrysChip linearizes DNA in NanoChannels

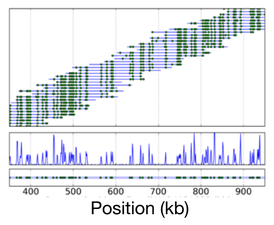


4) Irys automates imaging of single molecules in NanoChannels



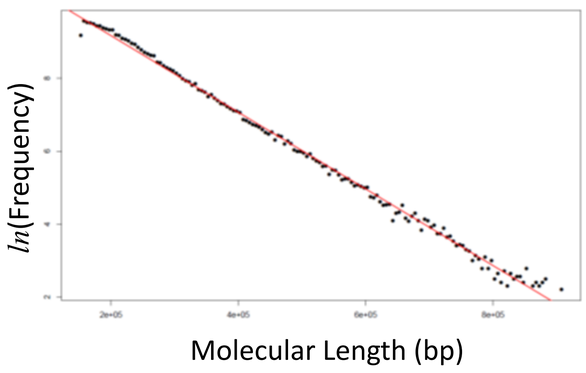
5) Molecules and labels detected in images by instrument software

6) IrysView software assembles genome maps

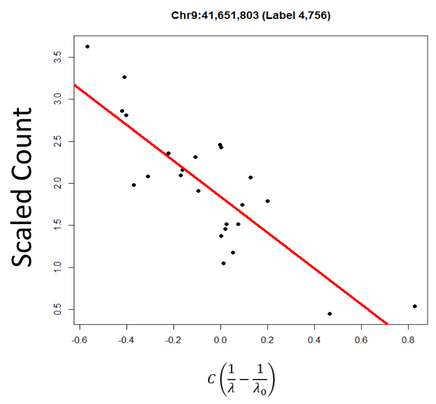


Evaluation of Sample-Specific and Reference-Specific Model Parameters

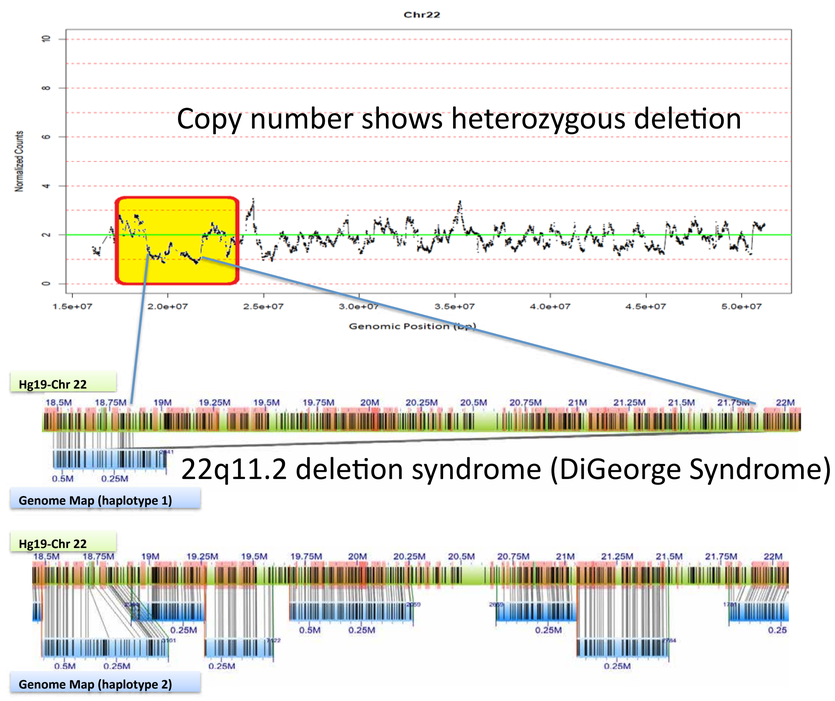
Average molecule length λ



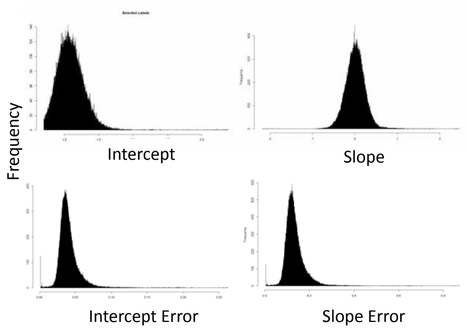
Intercept *I* and slope *S* per label



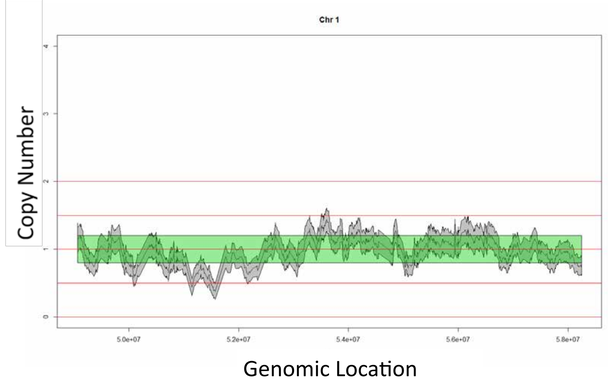
Heterozygous Deletion Detected via Copy Number



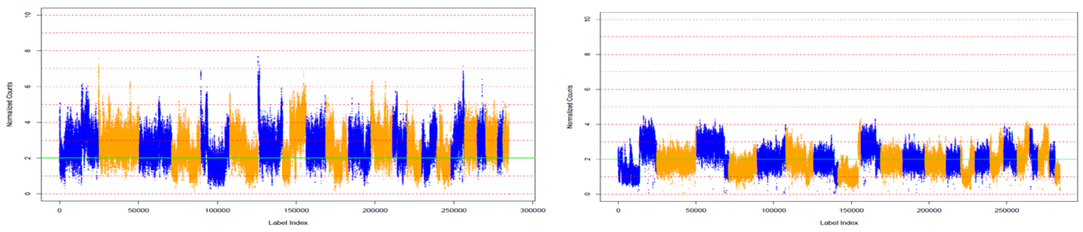
Model Parameter Values



Copy Number Profile & Uncertainty



Copy Number Profiles of Two Cancer Samples



The distribution of molecular lengths exponentially decays within the range from 150 to 800 kb. The average molecular length is therefore evaluated as the reciprocal slope of the linear regression between the molecular length and the logarithm of the frequency of molecules. Reciprocal average molecular lengths λ, measured for 24 euploid samples, along with the per-label raw depth counts Q (scaled with respect to sample-specific total autosomal counts), represent inputs for the set of linear regressions, yielding ~3.4x105 pairs of label-specific model coefficients (I, S). The per-label copy number values L are evaluated as follows:

$$L = \frac{Q}{I + SC\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)}$$

where C is a constant and λ_0 is the median of the average molecular length measured in the training set. The error in L is evaluated by second-order Taylor expansion:

$$(\delta L)^2 = \left(\frac{\partial L}{\partial Q}\right)^2 (\delta Q)^2 + \left(\frac{\partial L}{\partial I}\right)^2 (\delta I)^2 + \left(\frac{\partial L}{\partial S}\right)^2 (\delta S)^2 + \left(\frac{\partial L}{\partial \lambda}\right)^2 (\delta \lambda)^2$$

The derivatives are obtained straightforwardly and combined with uncertainty estimates reported by linear regressions, disregarding correlation and assuming Poisson distribution for raw counts. The resulting label-specific error estimates are used as weights in numerical processing and as acceptance criteria for visualization.

Conclusions

Long DNA molecules, recorded and quantified on the Irys platform, enable two-dimensional analyses of genome rearrangements. While the assembly yields an accurate geometric (horizontal) representation of the proband's genome, the coverage depth provides the copy number (vertical) information. By correlating the observed depth at a given genomic location with the sample-specific descriptor in a population of samples, we were able to reduce the relative error in per-label counts from ~40% to the stochastic limit, equal to the reciprocal square root of the average raw count per label. Model parameters are trained on population data and applied to predict copy number values in test samples. The resulting copy number breakpoints have resolution of ~10 kb, allowing comparison with assembled genome maps and structural interpretation of observed copy number variations. The wide dynamic range and the ability to estimate absolute (rather than relative) copy numbers renders this approach particularly suitable for analyses of complex genomes, containing amplified and/or rearranged regions. Applications to detection of congenital defects and analyses of cancer samples demonstrate the model's potential diagnostic utility.