Resizing N-Base Gaps in the Human Reference Genome

Željko Džakula, Alex Hastie, Andy Wing Chun Pang, Ernest Lam, Warren Andrews, Zhanyang Zhu, Xiang Zhou, Tiffany Y. Liang, Thomas Anantharaman, Han Cao*
BioNano Genomics, Research and Development, San Diego, CA
*Corresponding author

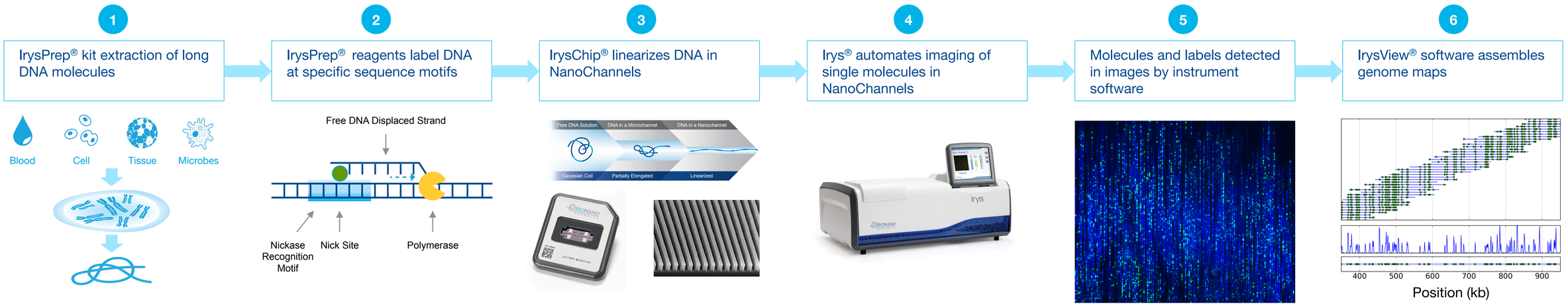
Abstract

With the Irys® System's next-generation mapping platform and NanoChannel technology, we have collected whole genome mapping data on an ensemble comprising of over 30 euploid human genomes. We then used the resulting de novo assembled genome maps to estimate the sizes of the N-base gaps in the human reference. The ultra-long DNA molecules (ranging from 150 kb to 2.5 Mb) enabled us to bridge a large fraction of N-base gaps present in the human reference and to accurately measure their widths. The corrected gap widths were evaluated by averaging the corresponding segment lengths observed in the euploid assemblies. The arbitrary sizes of N-base gaps currently present in the human reference significantly differ from our ensemble-based estimates, in sharp contrast with the high degree of consistency among the de novo assembled genome maps originating from different individuals. Our assemblies also suggest existence of an internal structure within a subset of N-base gaps. Using genome mapping data to evaluate this subset of gaps, we were able to accurately place sequence motifs inside a portion of the human genome previously considered "The Dark Matter." In addition, we analyzed reference regions surrounding N-base gaps that disagreed with our euploid assemblies. We are proposing an alternative structure for those regions, consistent with our observations both in terms of the number and the locations of the relevant sequence motifs. Finally, our genome maps reveal diversity among the individuals included in the ensemble, suggesting presence of haplotypes both inside and in the vicinity of the N-base gaps.

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 Irys 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 whole genome. The resulting order and orientation of sequence elements in the map can be used for anchoring next-generation sequencing contigs and structural variation detection. Here, we leverage the collective information extracted from a population of assembled euploid human genomes to correct gaps and errors in the human reference.

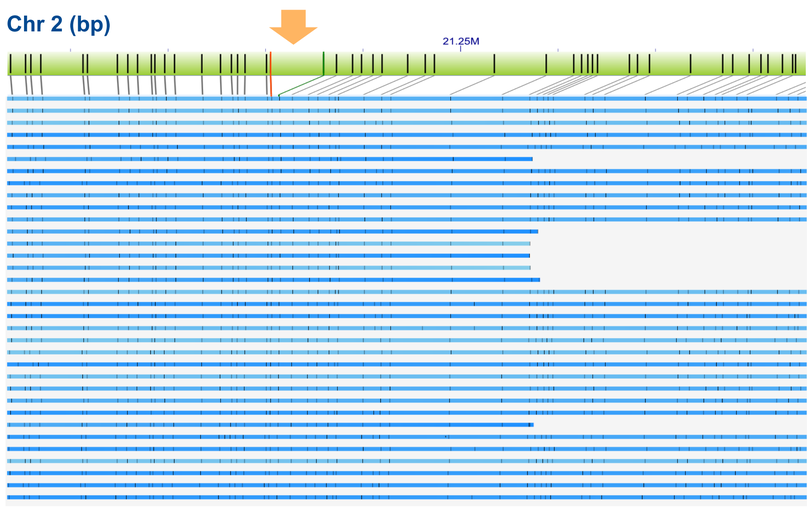
Methods



(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.

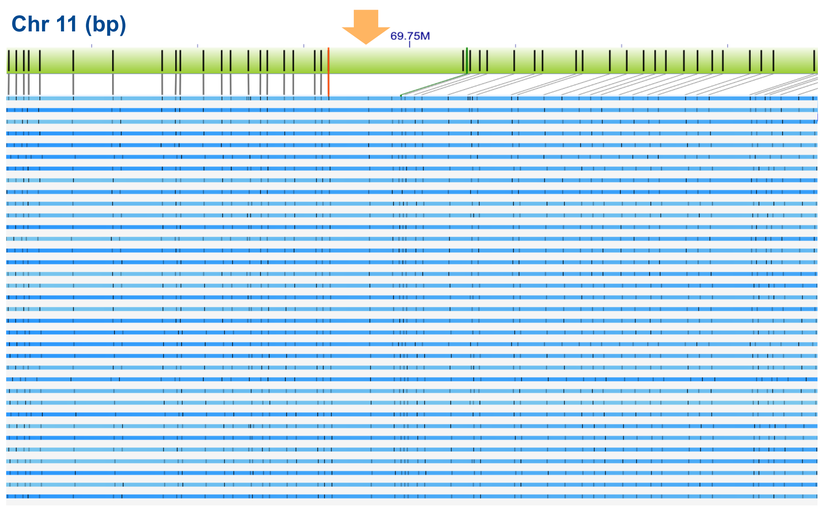
SIMPLE GAP SIZE REDUCTION

Number of gaps corrected	21
Starting total N-base length (bp)	1,340,200
Reduction of N-base length (bp)	1,190,347
Percent N-base length reduction	88.82%
Remaining N-base length (bp)	149,853
Percent of remaining N-bases	11.18%
Newly introduced labels	0

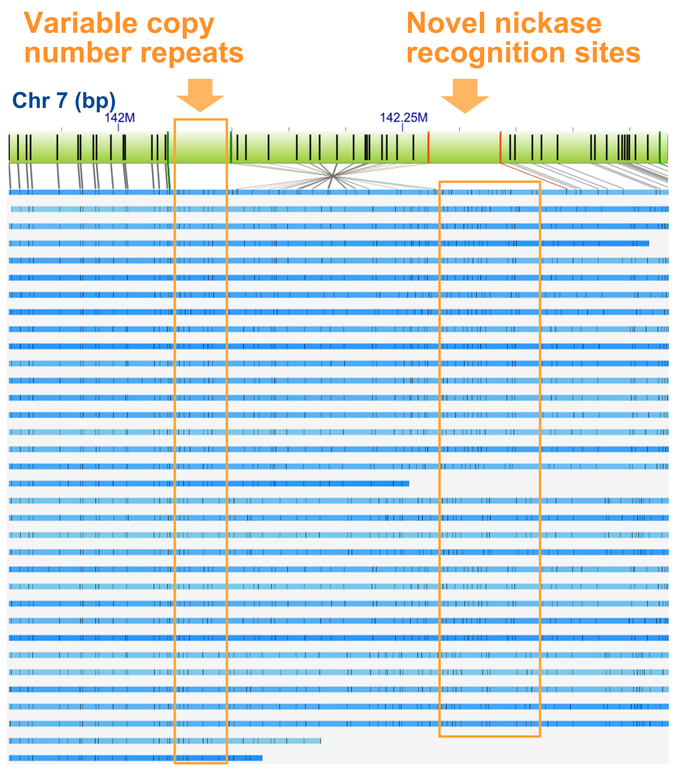


N-GAPS WITH NOVEL LABELS

Number of gaps corrected	18
Starting total N-base length (bp)	1,111,000
Reduction of N-base length (bp)	144,594*
Percent N-base length reduction	87.0%*
Remaining N-base length (bp)	217,178*
Percent of remaining N-bases	13.0%*
Newly introduced labels	25



REORIENTED REGIONS



GRCh38 validates the correction of hg19.

Conclusions

In addition to accurately measuring hg19 and GRCh38 N-base gap sizes on an ensemble of euploid human samples, we also identified novel Nt.BspQI nickase recognition site locations within the N-base gaps. Assembled genome maps from multiple human assemblies consistently indicate that orientations of several hg19 regions are wrong. All incorrectly oriented regions are flanked by N-base gaps. Several N-base gaps involve novel repetitive label patterns with variable copy numbers. The benefits of the corrected set of Nt.BspQI recognition site positions include improved structural variation detection accuracy and alignments of BioNano single molecules and genome maps. For more information about next-generation mapping, also see Posters #3118T, #2721W, #1632F and #2496F.

Reference

1. Cao, H., et al. Rapid detection of structural variation in a human genome using NanoChannel based genome mapping technology. GigaScience (2014); 3 (December 2014): 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.