Analysis of Biological Functions of Long Tandem Repeats in a Hummingbird Genome Using Next-Generation Mapping Technology

S Chan[1], E Jarvis[2], A Hastie[1], H Cao[1]

[1]BioNano Genomics, San Diego, CA, USA; [2]Duke University, Durham, NC, USA

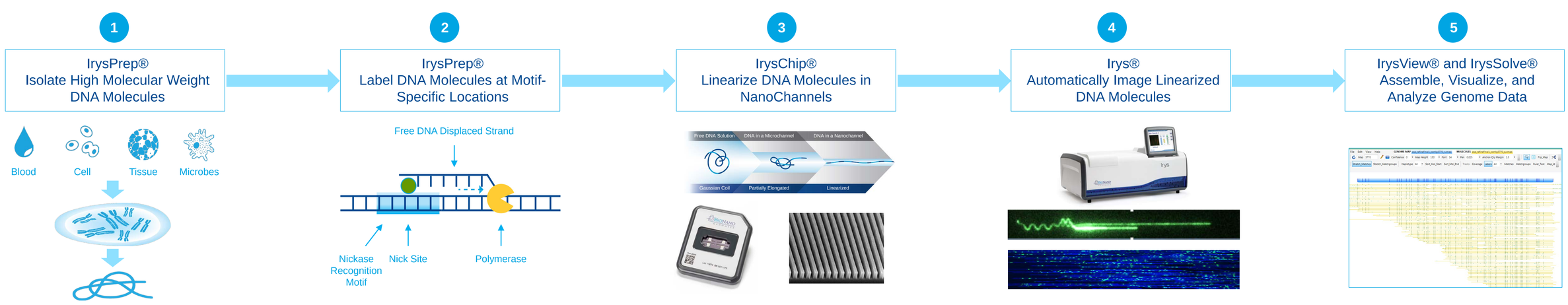
Abstract

Many plant and animal genomes are difficult to assemble because of the vast amount of long tandem repeat motifs, sometimes spanning several hundred kilobase pairs to multiple megabase pairs. Although repeat motifs can be identified and the amount of repeat material can be approximated by conventional sequencing technologies, it is difficult to assemble them into long contigs, so the exact locations and copy number of these repeats remain elusive, especially when the unit length exceeds the read lengths. Without knowing the genomic context or amount of repetitive material, it is impossible to attach biological relevance to them.

The BioNano Genomics Irys® System's next-generation mapping (NGM) technology enables imaging of intact megabase-scale molecules of DNA, so repeats can be spanned and properly placed in the genome assembly. By anchoring next-generation sequencing (NGS) contigs to genome maps, we can determine the sequence – and potentially the biological function – of the repetitive material.

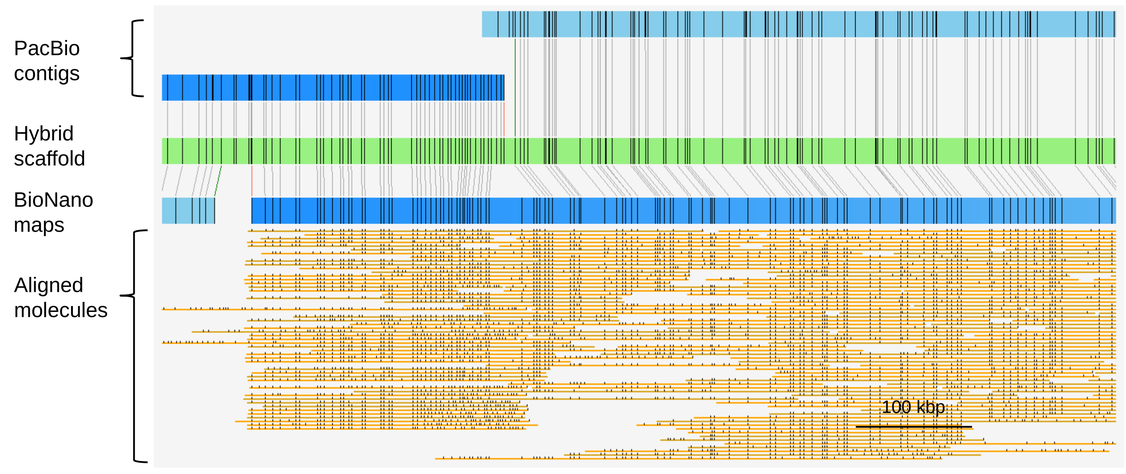
We present an analysis of repetitive regions in Anna's hummingbird (*Calypte anna*). Using the Irys System, we *de novo* assembled the genome, including repeat arrays intractable to current NGS methods, identified tandem repeats in BioNano maps using a novel algorithm, anchored Illumina and PacBio contigs to those maps, and investigated sequences within and neighboring those regions. We found a 60 to 120 kilobase pair (kbp) repeat contained within a muscle skeletal receptor tyrosine kinase gene, and a 106 to123 kbp repeat adjacent to a MAP/microtubule affinity-regulating kinase gene, which suggests that high copy number may correlate with increased muscle activity and glucose metabolism. A public database search suggests that these repetitive elements are hummingbird-specific.

Methods



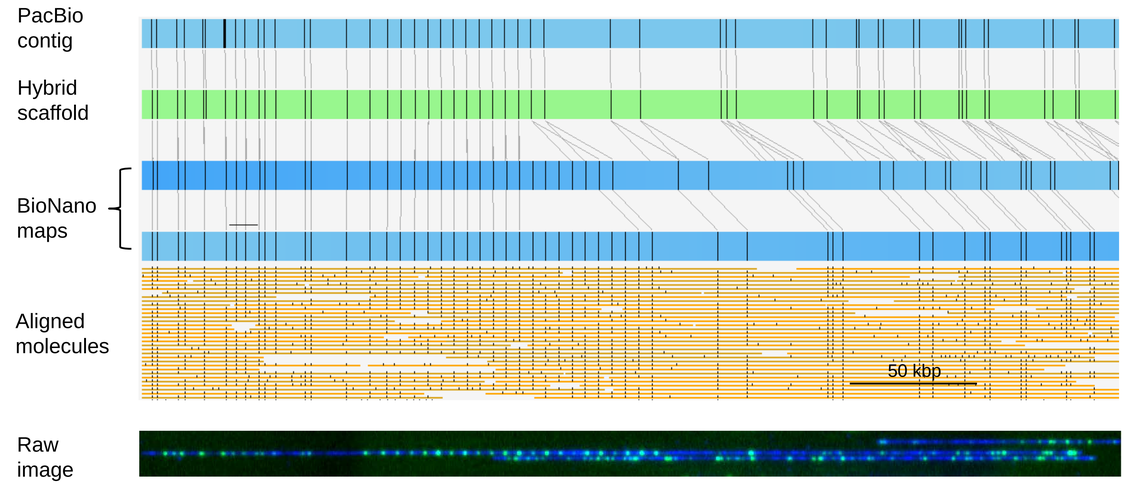
Scaffolding PacBio Contigs to BioNano Genome Maps to Span MuSK Gene

PacBio contigs and *de novo* assembled BioNano genome maps were combined to form a hybrid scaffold (green bar). The putative tandem repeat shown here, composed of 60 kbp of labeled and 60 kbp of unlabeled repeat material, was completely spanned by single molecules (beige lines), and assembled into a consensus genome map (bottom blue bar). This map was used to merge two PacBio contigs (*in silico* digest maps, top blue bars). Sequence data from inside each repeat unit align with both introns and exons of muscle skeletal receptor tyrosine kinase (MuSK) genes found in other birds. MuSKs are involved in the maintenance of neuromuscular junctions, and mutations in these genes in humans is known to cause several muscular diseases. The highest scoring alignment rates for the entire repeat array were those of two sabrewings (a type of hummingbird), suggesting that the high copy number is hummingbird-specific. Higher copy number may result in upregulation of the gene (or catalytic subunit(s)), corresponding to these birds' higher demand for muscle maintenance during flight.



De novo Assembly of Multiple Alleles and Their Correct Copy Numbers in MARK1 Gene

As shown in the example, the BioNano maps can also be used to resolve haplotypes. This tandem repeat, composed of 5.7 kbp repeat motifs, has been assembled in two separate BioNano maps, both of which have strong single-molecule support (shown here for only one of the maps). These are the two different alleles inherited from the parents. The BioNano maps have 19 and 22 sites, spanning 106 kbp and 123 kbp, respectively, while the *in silico* digested map of the PacBio contig has only 14 labels spanning 76.1 kbp (PacBio copy number was used in the hybrid assembly but may be revised). Sequence data aligns to a predicted MAP/microtubule affinity-regulating kinase (MARK1) gene. MARK1 belongs to the AMPK-related family of kinases, which are involved in the regulation of dynamic biological functions, including glucose and energy homeostasis. The high copy number may correlate with hummingbirds' high metabolism and ability to quickly convert food into fuel.



Conclusions

Irys technology complements NGS technologies by improving genome assembly contiguity, correcting assembly errors, and helping to resolve haplotypes, all of which are difficult (sometimes impossible) in repetitive regions. A well-assembled repeat array, along with its correctly placed flanking sequences, can be further investigated for genetic and biological function. This strategy of investigating repetitive material by combining the strengths of Irys and NGS, demonstrated here with the hummingbird genome, can be extended to all organisms and lead to a deeper understanding of the biological role of tandem repeats. See also Posters: P0207, P2078, P0236, and P1272.

Reference

- Lam, E.T., et al. Genome mapping on NanoChannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
- Sun, C., et al. Inactivation of MARK4, an AMP-activated Protein Kinase (AMPK)-related Kinase, Leads to Insulin Hypersensitivity and Resistance to Diet-induced Obesity. *The Journal of Biological Chemistry* (2012); 287:38305-35315
- Welch Jr., K.C. and Suarez, R.K. Oxidation rate and turnover of ingested sugar in hovering Anna's (*Calypte anna*) and rufous (*Selasphorus rufus*) hummingbirds. *The Journal of Experimental Biology* (2007); 210:2154-2162