Leveraging Genome Mapping in Nanochannel Arrays and NGS for a Better Human *De Novo* Sequence Assembly

H. Dai[1], A. Pang[1], W. Stedman[1], T. Anantharaman[1], A. Hastie[1], P.Y. Kwok[3], A. Ummat[2], E. Schadt[2], R. Sebra[2], B.A. Bashir[2], H Cao[1]
[1]BioNano Genomics, San Diego, CA; [2]Mount Sinai School of Medicine, New York, NY;
[3] University of California, San Francisco, San Francisco, CA

Abstract

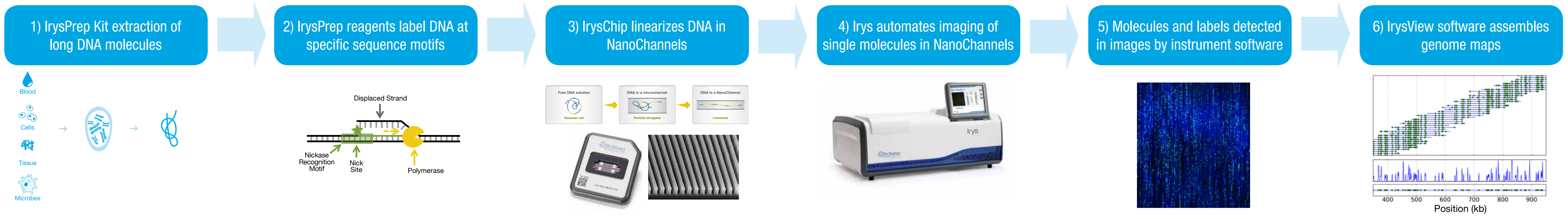
Irys genome mapping technology provides direct analysis of extremely long genomic DNA (up to multi-megabases) without amplification. *De novo* assembly of these single molecules yields high-fidelity contiguous map information across long ranges. Its advantage over all other genome assembly methods is particularly dramatic in highly repetitive regions. Genome maps thus greatly complement assemblies using relatively short second- and third-generation sequencing reads. We have constructed genome maps of human NA12878 (cell line derived from the daughter in the CEU trio) which resulted in a consensus assembly measuring 2.9 Gb and with an N50 of 4.6 Mb. With Pacific Biosciences sequence from the same cell line, we also created a sequence-based assembly with N50 length of 930 kb in parallel. By combining data from these two technologies with a custom-designed hybrid scaffolding pipeline, we were able to generate an assembly having scaffold N50 length of greater than 30 Mb covering more than 2.7 Gb of the human genome. At the same time, we were able to identify potential misassembles in the sequence assembly as well as in the genome maps by reviewing the inconsistencies between these two complementary technologies. The hybrid scaffolds also discovered additional long range structural variations not identified in the sequence assembly.

Background

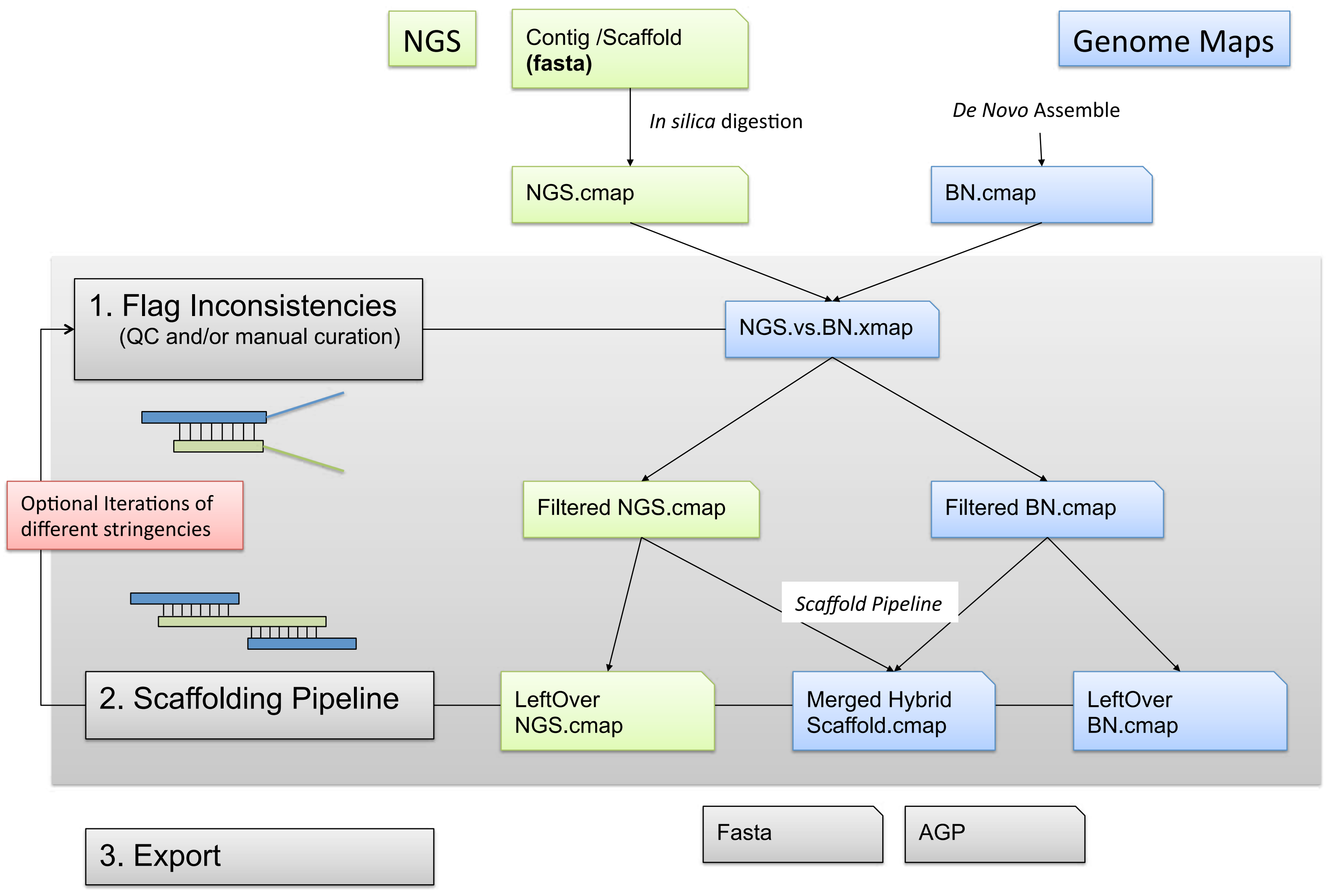
Generating high quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. Instead, Irys technology provides direct visualization of long DNA molecules in their native state, 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

Hybrid Scaffolding Flow Chart


NGS
Contig /Scaffold **(fasta)**
Genome Maps
In silica digestion
De Novo Assemble
NGS.cmap
BN.cmap
1. Flag Inconsistencies (QC and/or manual curation)
NGS.vs.BN.xmap
Optional Iterations of different stringencies
Filtered NGS.cmap
Filtered BN.cmap
Scaffold Pipeline
2. Scaffolding Pipeline
LeftOver NGS.cmap
Merged Hybrid Scaffold.cmap
LeftOver BN.cmap
3. Export
Fasta
AGP

Flow chart for the hybrid assembly strategy.

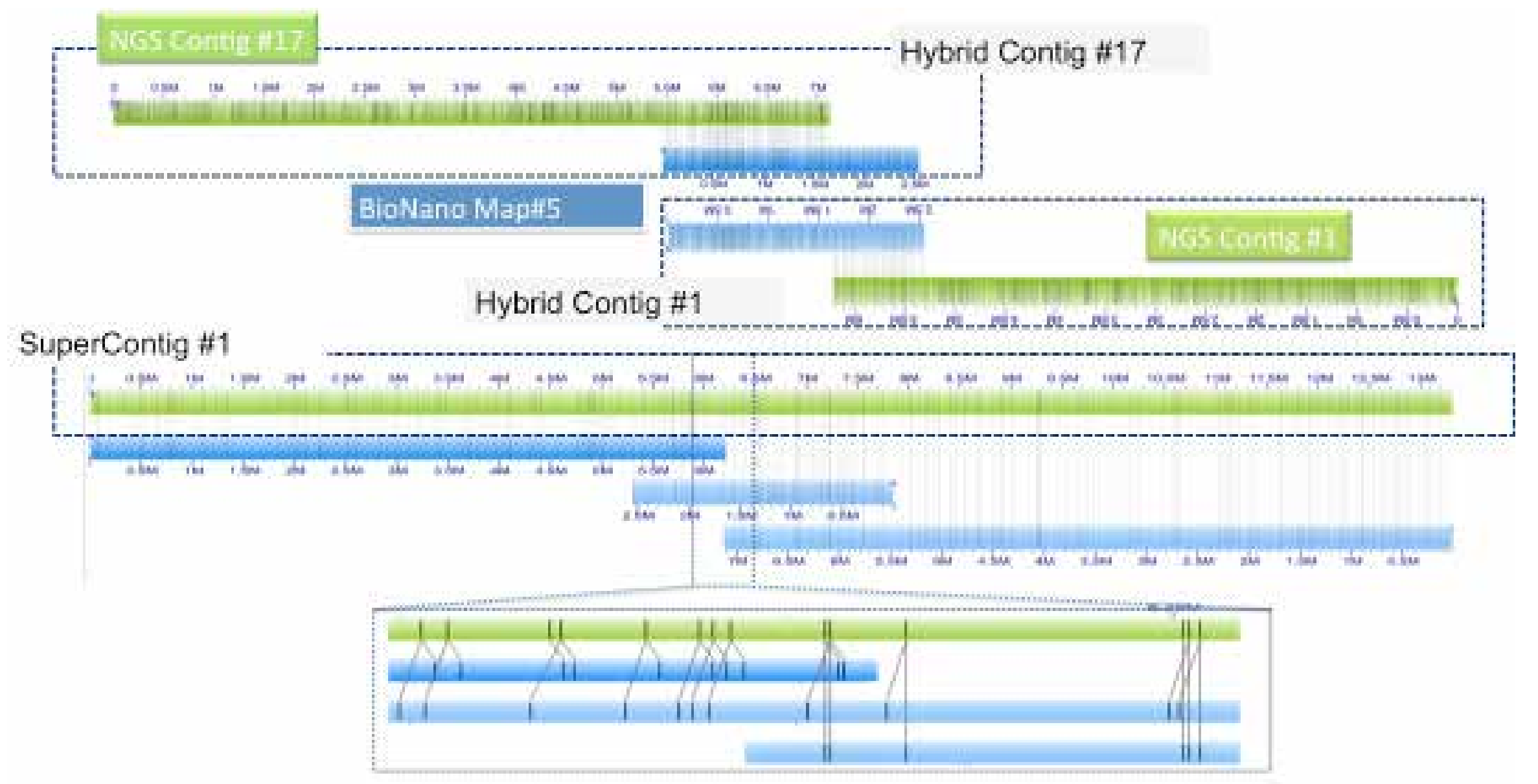
NGS & Genome Map Hybrid Assembly

Step 1: Merge between NGS/BioNano contigs to generate hybrid contigs, keep merge cascade between hybrid contigs and NGS/BioNano contigs but not inside groups (NGS vs. NGS, BioNano vs BioNano and Hybrid vs Hybrid)

Step 2: Merge between hybrids to generate super scaffolds

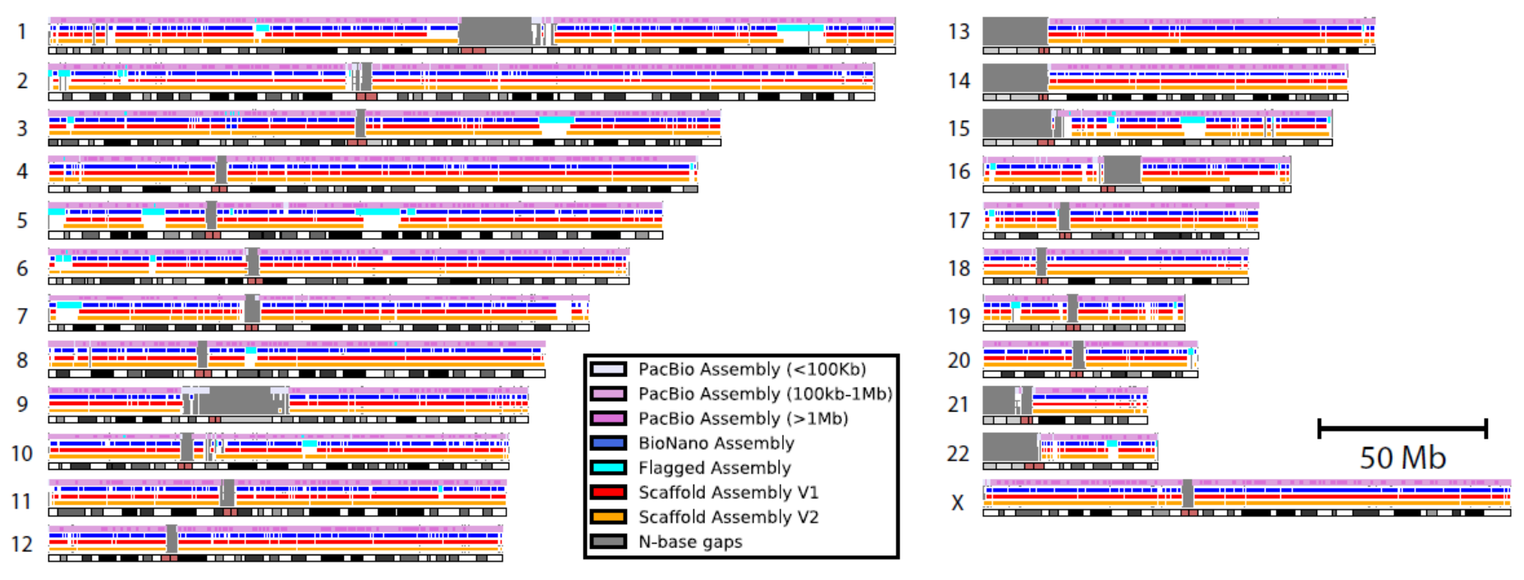
Step 3: Extend hybrid with BioNano extend/merge pipeline

Step 4: Align NGS contigs to super contigs



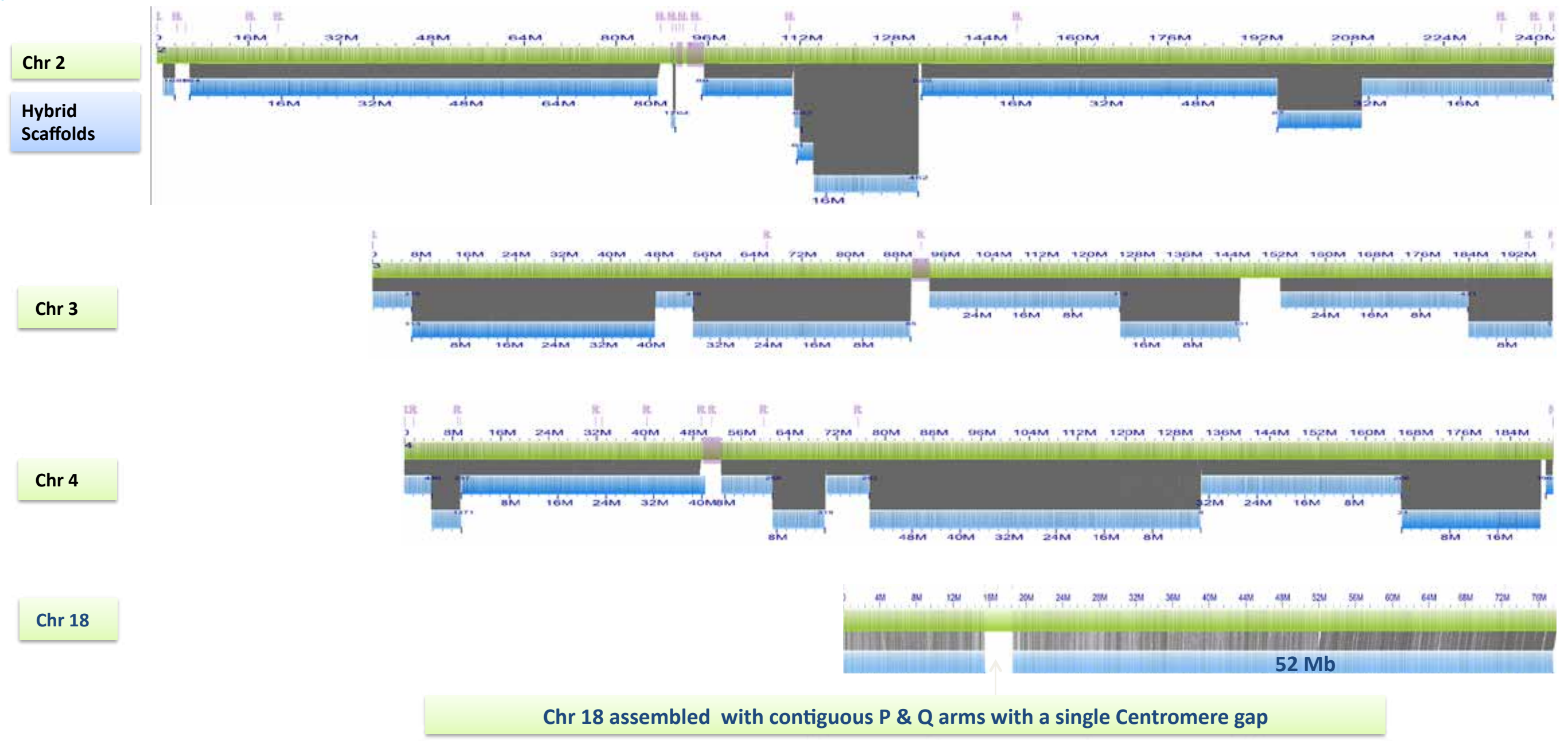
Coverage & Assembly Size for Various Assemblies

Assembled contigs, genome maps and scaffolds are ordered relative to hg19. Moving outwards from the reference: PacBio sequence contigs (color gradient from light purple (<100kb) to dark purple (>1 Mb)), Bionano genome maps (blue), scaffold V1 (red) and scaffold V2 (orange). Also, possible inconsistencies identified between sequence and Bionano data are labeled in cyan. The ideogram is plotted at the bottom of each chromosome, with centromeres highlighted in red. Also, N-base gaps in the reference genome are shaded in grey in the background of all contigs and scaffolds.



Towards True Contiguity

Examples of Chromosomal Level Hybrid Assemblies of Human Chromosomes



Chromosomal alignment views of hybrid scaffolds (blue) to hg19 reference (green). With the exception of centromeres or a few loci, the chromosomes are fully covered by hybrid scaffolds. In fact, we see that each chromosome arm is usually spanned by just a handful of hybrid scaffolds, thus exemplifying the long contiguity of the data. Most notably, our scaffolding pipeline yields single scaffolds in both arms of chromosome 18.

Hybrid Assembly

Combining NGS Contigs and Genome Maps

	# Contigs	N50
NGS Contigs	17,199	930 Kb
Selected NGS Contigs	3,895	**1.1 Mb**
BioNano	1,003	4.6 Mb
Hybrid Scaffolds	182	**34.5 Mb**

Overview of the results of the co-assembly show that the N50 values are **30X** better than PacBio alone and **6X** better than genome map alone.

Conclusions

Our data of this human sample NA12878 illustrates that pairing technologies can enable superior results than can be achieved with any single technology. The extremely long BioNano genome maps facilitate the correct ordering and orientation of Pacific Biosciences contigs. In fact, we achieved a hybrid assembly that has a larger N50 value than any shot-gun sequencing project of the human genome to date. Furthermore, our pipeline can generate outputs in FASTA format, and so users can zoom in to base-pair resolution. Therefore, our hybrid data can provide insights from large structural level down to nucleotide level. We anticipate that future fully-integrated co-assembly of BioNano's long molecules with sequence reads can further improve the contiguity and accuracy of the resulting assembly, thus enabling an even deeper understanding of the biology of genomes.

References

1) Ummat, A. et al. Assembly and Diploid Architecture of an Individual Human Genome via Single Molecule Technologies. In preparations.
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
4) Das, S. K., et al. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Research (2010); 38: 8
5) Xiao, M et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.