ASSEMBLING HIGH QUALITY HUMAN GENOMES: GOING BEYOND THE '$1,000 GENOME'

Case Study | October 2015

Scientists from WashU, Macrogen, and Mount Sinai are using long-read sequencing with single-molecule, next-generation genome mapping to create gold-quality de novo assemblies of human genomes. Unbiased de novo assembled genomes also highlight the substantial amount of structural variation unique to individuals and populations, which cannot be accessed by short-read technologies that use a reference-based re-sequencing approach.

Even as the human genetics community generates genome sequencing data at unprecedented pace, researchers still struggle to pinpoint the genetic factors underlying many diseases. A significant limitation, scientists have discovered, is that re-sequencing solely with short-read data lack full coverage of the human genome[1] and fail to resolve important genetic variations such as disease-causing structural variants in ALS, autism, and cancer[2]. To overcome this problem, leaders in the sequencing field are embracing new long-read technologies that provide the most complete view yet of individual human genomes.

These pioneers use Single Molecule, Real-Time (SMRT®) Sequencing together with single molecule, Next Generation Mapping (NGM) from BioNano Genomics' Irys® System to produce the most contiguous, accurate, and comprehensive ***de novo*** human genome assemblies[3] on the market today. They have demonstrated that the richness of information gathered from this technology combination results in high-quality assemblies that can be produced for any individual or population. This union of complementary technologies represents a critical advance in enabling the community to move beyond resequencing methods, an approach limited by its reliance on a single reference genome and subject to a variety of known biases[4].

SMRT Sequencing from Pacific Biosciences generates industry-leading read lengths and the highest consensus accuracy, making it possible for scientists to produce high-quality ***de novo*** assemblies. Scientists have used PacBio® data to generate high-quality haploid assemblies of hydatidiform mole cell lines as well as several diploid assemblies of human genomes.

NGM from BioNano Genomics produces an accurate representation of very long single molecules ranging in length from 150,000 base pairs to more than 1 million base pairs. Treated with nicking enzymes and digitized as they flow through the IrysChip®, these individual molecules yield unadulterated, long-range information without PCR bias, allowing users to generate comprehensive and accurate ***de novo*** assemblies.

Assembly / Scaffold Summary Statistics

	CHM1	CHM13
PacBio *de novo* Assembly Contig N50	27.94 MB	12.98 MB
BioNano + PacBio Scaffold N50	50.57 MB	22.45 MB

> "With PacBio we get much longer N50 contigs than we ever achieved before and by combining it with BioNano data we almost double that."

The pairing of these orthogonal technologies produce a more complete and higher-quality genome assembly than either solution offers on its own. Scientists report that PacBio data provides the most contiguous assembly for use with a BioNano map, while BioNano maps enhance the long-range scaffolding and accuracy of PacBio assemblies. Research teams have shown repeatedly that ***de novo*** human assemblies generated by integrating these two data types deliver the most comprehensive view of the genome, detailing structural variants and other complex genomic elements that are systematically missed by short-read resequencing methods.

Here, three leading scientists who have paired PacBio and BioNano technologies discuss their efforts to improve existing genome references and to produce entirely new high-quality ***de novo*** human genome assemblies .

The Ultimate Reference

Tina Graves-Lindsay spends more time than most thinking about the perfect human genome. As leader of the Reference Genomes Group at the McDonnell Genome Institute at Washington University and a member of the Genome Reference Consortium (GRC), she focuses on building tools that will serve the community for years to come. "Our goal is to create assemblies that could be

Tina Graves-Lindsay, Leader of the Reference Genomes Group at the McDonnell Genome Institute at Washington University

used as a reference," she says.

To that end, her team has been working hard to create the highest-quality human genome assemblies. They get superior results using PacBio long-read sequencing together with BioNano Genomics whole genome mapping. So far, this method has helped them produce two haploid human genomes based on hydatidiform mole cell lines (CHM1 and CHM13) and one diploid assembly of the human genome. Another diploid genome is underway, and Graves-Lindsay says plans are to sequence at least three more of these. These initiatives have been funded by an NIH grant aimed at adding highly accurate information to the existing human reference genome.

In the hands of Graves-Lindsay, the pairing of PacBio and BioNano data is particularly useful for increasing contiguity of genome assemblies. Her team assembles the human genome using SMRT Sequencing data, and independently creates a whole genome map with the BioNano Irys System. Once both data sets are complete, the scientists integrate and compare them. "BioNano helps inform us of contigs that need to join, or places that are potentially incorrect in the PacBio assembly," Graves-Lindsay says. "In the end you get a much better final product."

Adding an Irys genome map to a SMRT Sequencing assembly routinely boosts contiguity, and both technologies significantly outperform what the team was generating with short-read data. "Using PacBio, we're getting much higher N50 contig lengths than we ever achieved before with *de novo* assembly," Graves-Lindsay says. "We almost double that in most of the examples I've looked at so far by using the BioNano in addition to the assembly."

Based on her experience with *de novo* sequencing for reference-grade genomes, Graves-Lindsay says she hopes the community is becoming aware "that you can get so much more with a *de novo* PacBio assembly as opposed to a *de novo* Illumina assembly." While her focus is on creating reference tools for the community, any *de novo* assembly of a human genome would still benefit from the combination of long-read sequencing and genome mapping, compared to the gaps and partial transcripts characteristic of short-read assemblies. "These highly contiguous assemblies have much greater utility when doing genomic comparisons," she adds.

Graves-Lindsay believes that the community won't stop after achieving a handful of reference-quality human genomes. "The more we sequence, the more diversity we find," she says. "There's going to be an ongoing need to generate high-quality *de novo* assemblies."

But for now, she'll focus on producing the highest-quality reference genome, for which she will apply the combination of PacBio sequencing and BioNano genome mapping. "Our ultimate goal is not a high volume of low-quality assemblies, but rather highly accurate and contiguous assemblies that provide the most utility for genomic studies," Graves-Lindsay says.

Population Resource

In Korea, scientists used PacBio sequencing and BioNano Genomics to generate the world's first Asian diploid *de novo* assembly based on long-read data. Jeong-Sun Seo, chairman of Macrogen, Inc., and professor at Seoul National University College of Medicine, helped lead this effort, which was aimed at producing a population-specific reference genome to shed more light on genetic variation among Asians.

Leading an effort to produce a population-specific reference genome to shed more light on genetic variation among Asians

Korean scientists have developed a number of resources for the population, including a *de novo* assembly and a copy number variation map. Still, Seo realized they would need what he calls "a medical-grade genome" to have the biggest impact. Mapping short reads to an incomplete reference, he says, is simply not practical. The team used multiple platforms, including SMRT Sequencing and BioNano genome mapping, to produce a new diploid *de novo* assembly for a Korean genome, where N50 supercontig length was ~22 Mb. In addition, BACs were used not only to validate their assembly, but also to provide a haplotype-resolved genome. "The project was initiated with the goal of providing a haplotype-resolved reference

Professor Jeong-Sun Seo, Chairman of Macrogen, Inc., and Professor at Seoul National University College of Medicine

genome that is most representative of the Asian population," Seo says, noting that the resulting assembly will be critical for enabling precision medicine in the Asian population.

The quality of the resulting assembly was so good — including phased haplotype blocks with N50 lengths of ~10 Mb when combined with BAC clones and other single molecule technologies — that Macrogen scientists used it to fill or reduce gaps in the official human genome reference, GRCh38. Much of that new sequence was highly repetitive, or consisted of long insertions and deletions, regions known to stymie short-read sequencers but which can be fully characterized with long-read sequence data.

"The combination of PacBio sequencing and BioNano Genomics Next Generation Mapping yields results that are greater than the sum of its parts," Seo says. "PacBio was pivotal in providing the necessary read length for *de novo* assembly and for investigating the diploid structure of the AK1 genome while BioNano maps were essential in increasing the contiguity of our assembly and in correcting misassembled regions. Together, the two technologies open previously inaccessible regions of the genome."

By comparing the Korean assembly to GRCh38, the scientists were able to find population-specific novel structural variation. For example, a structural variant in the MUC3A and PADI4 gene is associated with disease among Koreans but not among people of European ancestry[5].

Going forward, Seo says this medical-grade genome will maximize the value of other initiatives, such as a program to sequence as many as 10,000 people of Asian ancestry to catalog genetic variation. Those new sequences will be far more useful now that the team can compare them to a high-quality reference assembly produced with BACs, PacBio sequencing, and BioNano next-generation mapping technologies.

The Complete Picture

At the Icahn Institute for Genomics and Multiscale Biology, scientists have used the technology combination to produce "the most contiguous clone-free human genome assembly to date," as they described it in a *Nature Methods* paper[3].

Eric Schadt, founding director of the Icahn Institute, says his team aimed to generate a reference-grade human genome assembly in part to determine which technologies are best suited for that goal. "To get as complete a genome as possible, clearly no single technology gives you that on its own," he says. "You get increasing resolution by integrating and combining these different technologies."

Scientists from the Icahn Institute worked with collaborators at a number of organizations, including the National Institute of Standards and Technology, to evaluate various technology options for the shared goal of creating a high-quality human genome assembly. They used the well-characterized NA12878 genome to provide a strong foundation for their effort.

> "If your aim is to uncover the most variation you can to explain human disease or human biology, your only solution is to employ these more advanced technologies."

"Our analysis of the NA12878 genome shows that combining complementary technologies yields results that are superior to those from any single technology," Ali Bashir and team report in the *Nature Methods* paper. "Long contigs from SMRT Sequencing facilitate unambiguous mapping to genome maps ... Analogously, although long reads elucidate [structural variants] far better than short reads and provide breakpoint-level precision, some events contain repeat lengths that only genome maps can accurately resolve." The team compared the resulting data to the human reference genome and detected discrepancies that were resolved through this multi-platform approach.

Eric Schadt, Founding Director of the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai

The scientists note that a heavy reliance on short-read sequencing has hindered *de novo* analysis of human genomes with a proliferation of short scaffolds that cannot span large genomic elements. "Short-read technologies are wholly incapable of addressing heterochromatin DNA regions — centromeres and telomeres — because of their low-complexity, repeat-rich nature," Schadt says. Even in euchromatin, "there are lots of structural variations, extensive repeat regions, long tandem repeats, and so on that are not adequately addressed with short-read technologies. We're missing a significant proportion of the genome."

Missing those features is detrimental to a comprehensive understanding of the human genome and disease causation, Schadt says, pointing to the repeat expansion driving Fragile X syndrome and the long tandem repeats in mucin genes as two examples. "Without question we're going to find that those structural features are incredibly important for understanding disease biology," he says. "If your aim is to uncover the most variation you can to explain human disease or human biology,

your only solution is to employ these more advanced technologies."

At the Icahn Institute, scientists have deployed long-read sequencing from PacBio as well as genome mapping from BioNano Genomics. While the pairing of these two technologies represents an extra investment in time and resources, Schadt believes that costs will continue to drop and that this method will become a standard practice in genomics labs, enabling scientists to find transposable elements, segmental duplications, repetitive elements, and other structural variants far more reliably than they can with the typical short-read sequencing pipeline.

"Individually, the assemblies and genome maps markedly improve contiguity and completeness compared with *de novo* assemblies from clone-free, short-read shotgun sequencing data," the researchers note in the publication. "By combining the two platforms, we achieve scaffold N50 values greater than 28 Mb, improving the contiguity of the initial sequence assembly nearly 30-fold and of the initial genome map nearly 8-fold."

"The very significant improvement in the NA12878 genome when the technologies were combined is a really good indicator that we're not as far along as we think we are on understanding genomes," Schadt says.

This is far more than an academic exercise to get the best human assembly: for Schadt, it's all about finding the method that will make a difference in people's lives. "*De novo* assembly of genomes is going to be critically important to getting to this goal of precision medicine," he says. "Moving beyond reference-based mapping to *de novo* assembly of each individual's genome is really going to be where the action is."

References

1. Nature Genetics Editorial. (2015) Whole genome?, *Nature Genetics*, 47, 963
2. Baker M. (2012) Structural variation: the genome's hidden architecture, *Nature Methods* 9, 133–137
3. Pendleton M. et al. (2015) Assembly and Diploid Architecture of an Individual Human Genome via Single Molecule Technologies, *Nature Methods* 12, 780-786
4. Bustamante CD, Rasmussen M. (2015) Beyond the reference genome, *Nature Biotechnology* 33, 605–606
5. Lee HS et al. (2009) Lack of Association of Caucasian Rheumatoid Arthritis Susceptibility Loci in a Korean Population, *Arthritis Rheum.* 60(2): 364–371

www.pacb.com

Headquarters
1380 Willow Road
Menlo Park, CA 94125
United States
Phone: 1-650-521-8000

Singapore Office
20 Science Park Road
#01-22 TeleTech Park
Singapore 117674
Phone: +65 67785627

Customer Service
Phone: 1-877-920-PACB (7222) option 1
Fax: 1-650-618-2699
Email: orders@pacificbiosciences.com

Technical Support
Phone: 1-877-920-PACB (7222) option 2
Email: support@pacificbiosciences.com

PacBio Sequencing Providers
www.pacb.com/SMRTproviders

Sales Inquiries
North America:
nasales@pacificbiosciences.com

South America:
sasales@pacificbiosciences.com

Europe/Middle East/Africa:
europesales@pacificbiosciences.com

Asia Pacific:
apsales@pacificbiosciences.com

www.bionanogenomics.com

Headquarters
9640 Towne Center Drive, Suite 100
San Diego, CA 92121
United States
Phone: 1-858-888-7600

Customer Support
Phone: 1-858-888-7600
Email: support@bionanogenomics.com

Sales
Phone: 1-858-888-7600
Email: sales@bionanogenomics.com

Service Providers
Phone: 1-858-888-7600
Email: info@bionanogenomics.com

Distributors
www.bionanogenomics.com/about-us/distributors

PN: CS111-092415