DECODING THE UNSEQUENCEABLE

In case you're not aware, the human genome is not completely sequenced. Regions of repeats and hard to decode sequence continue to elude scientists. And then there is the challenge of genetic variation. Nathan Blow looks at new technologies and methods that could help map these and other difficult-to-read stretches of DNA.

A sequencing room at the Beijing Genomics Institute.

There are things in everyday life that we tend to take for granted. The world of life science methods is no different in many ways. Take DNA sequencing, which has become such a commonplace tool in the modern lab that researchers often assume any section of DNA can be decoded and analyzed once a sample is placed inside a next-generation sequencing machine. But it turns out that there are instances where the robust and reliable technique of DNA sequencing, even when coupled with the latest bioinformatics software, can run into problems that require extra attention and skill.

Genomes are primarily composed of long stretches of G, A, T, and C nucleotides in various arrangements. When these nucleotides are arranged in “normal” configurations (such as in the coding sequence of a gene), all is generally good, and sequencing is usually not difficult. The challenge starts when these arrangements become less diverse, when nucleotides start to repeat in large blocks or arrays, or when single nucleotides stretch on and on (a situation known as a homopolymer). This is the moment when even the most advance sequencing systems and bioinformatics assembly programs can run into a little trouble, leaving genome assemblies with holes.

“I don’t think there has been a single finished human genome on the planet yet,” says Deanna Church, Senior Director of Genomics and Content at Personalis, Inc., who has been involved with efforts to finish the human genome sequence for years now. “At least not one I have seen.”

Church, along with a dedicated collection of sequencing and genome assembly specialists, has been chipping away at the unsequenceable regions of the human genome since

the publication of the draft human genome more than a decade ago. Their efforts have resulted in new approaches and tools for DNA sequencing as well as unique insights into biology. But as they make more and more progress and more genomes are sequenced, another issue is beginning to come to light—what will the finished reference genome look like? How can vast amounts of genetic variation be captured in a single reference sequence? As it turns out, the biggest impact of sequencing the human genome might still be coming—and it will problem mean a shift in the way that we think about a “finished” reference genome.

Technical Limits

Current crop DNA sequencing platforms produce significantly higher outputs than ever before. But there is a tradeoff for that high throughput—these platforms also tend to produce much shorter sequence reads than the systems used in the past. And the trouble with shorter read lengths really can come to light when scientists are trying to assemble complex genomes from millions and millions of very short sequences.

Imagine a jigsaw puzzle with several thousand extremely small pieces, all having roughly the same shape and color. The more fragments of increasingly smaller size, the more challenging it becomes for someone to put the puzzle back together (the reason we avoid 5000 piece puzzles of the sky at the toy store). This is very similar to the challenge confronted by researchers and bioinformaticians as they try to take advantage of the throughput of today’s sequencing systems with their shorter read lengths.

It is for this reason, Church says, that when attempting to assemble large, complex genomes (such as those of mammals), it is critical to employ different sequencing techniques and technologies, as well as to use some non-sequence data at times.

One approach to solving the “small piece” problem is to generate a framework or a foundation upon which to place all your smaller pieces—think of it as being given the large borders around our puzzle. In this way, you have a defined starting point upon which to piece together the rest of the puzzle. When it comes to genome sequencing, this is where the emergence of new long-read sequencing systems might have the greatest impact.

Pacific Biosciences, located in Menlo Park, California, has commercialized a single-molecule sequencing system that offers very long reads of DNA fragments. In fact, the company's latest platform has been reported to produce single reads several thousand base pairs in length, considerably longer than the several hundred bases per read generated by the Illumina or Life Technologies short-read platforms. While the accuracy level and throughput are less than Illumina, data from the Pacific Biosciences system has been used as a “scaffold” for working with larger genomes, where longer reads guide the placement of the shorter read data from other systems. Such combination approaches have shown promise in recent years.

In 2012, a team of researchers led by Sergey Koren and Adam Phillippy from the National Biodefense Analysis and Countermeasure Center in Frederick, Maryland reported in the journal *Nature BioTechnology* (1) on the integration of high-fidelity short read sequences with multi-kilobase single molecule sequencing reads to correct errors and obtain de novo genome assemblies. To demonstrate the robustness of this approach, Koren, Phillippy, and their colleagues showed data for phage, prokaryote, and eukaryote genomes, as well as

Deanna Church has been working on finishing the sequencing of the human genome. Image courtesy of Deanna Church.

the genome of the parrot, which was previously unsequenced, at 99.9% base-calling accuracy. While assembly of larger genomes can be facilitated through the integration of multiple technologies, smaller genomes, such as those of bacteria, often do not require these additional steps. In fact, in 2013 in the pages of *Nature Methods* (2), another team, led by Jonas Korlach at Pacific Biosciences, reported the ability to assemble finished microbial genomes using Pacific Bioscience SMRT sequencing data alone.

Repeating, repeating, not reading

If you're looking for a particularly vexing region of the human genome to sequence, look no further than the centromere, the region that links sister chromatids. When it comes to sequencing centromere DNA, the stumbling block lies mainly in its composition—centromeres are composed of long repeating arrays DNA sequence. To make matters worse (or somewhat better depending on how you look at it), the sequences within those repeats are similar, but not identical, making ordering the arrays a difficult bioinformatics challenge.

“In the original human genome assembly, centromeres only appeared as a run of 3 million N's,” recalls Church, who described the “N's” as placeholders in advance of future finishing efforts. Following the completion of the human genome draft sequence in 2003, the Genome Reference Consortium (of which Church was a member) started working on decoding such difficult regions to finish the human genome and remove gaps in the sequence. Regions such as centromeres would send Church and her other finishing colleagues on unique sequencing excursions wherein creativity, modified sequencing approaches and enhanced informatics technologies would be the keys to success.

For centromeres, it would not be until 2014 that Karen Miga and her colleagues first reported on their efforts to decode

The Pacific Biosciences RSII instrument is a single-molecule sequencing platform capable of generating long DNA sequencing reads. Researchers are exploring the combination of short- and long-read technologies for de novo genome assembly.

a 3.8 million nucleotide stretch of centromeric DNA from a single individual genome in the journal *Genome Research* (3). Using whole genome sequencing reads from the previously published Venter genome, Miga and her co-workers first identified those sequence reads containing centromeric array repeat sequences. From there, the authors were able to construct models of the repeat numbers and their order within the genome. Note that the word “model” here instead of “reference sequence.” While this model of the centromere sequence should prove useful to researchers exploring centromere structure and function, it’s important to note that these sequences are not exact matches to the centromere sequences found in the Venter genome. What Miga and her colleagues created was a model that represents both the variants and the diversity of sequences within the centromere region. And while not perfect matches, it turns out that such models might just be the next new wave when it comes to describing reference genomes.

The new face of the reference genome

All of these specialized genome sequencing and finishing efforts started with a single purpose in mind—to generate a full human reference genome that could be used by

Stephen Quake's lab used microfluidic approaches to separate chromosomes for single cell haploid sequencing.

researchers studying human genetics and biology. But the concept of a single human reference genome brings up another question researchers are now wrestling with—whose reference genome should we be looking at? There exists a significant amount of genetic variation within the human population (single nucleotide polymorphisms, structural variations), so should we incorporate sequence variation into a reference genome to make annotations more useful to the scientific community? Is that even feasible? Some scientists think that in order to make this happen, we might need to rethink the concept of a reference genome.

Sequences are most often depicted in molecular biology as linear strings of A, G, C, and T nucleotides. And pairs of chromosomes are often broken down into a single sequence that can be used for comparisons. The trouble here is that there is more than a single copy of each chromosome in each individual, so how can you tell if two mutations in a gene are on the same allele or on different alleles when sequencing a diploid sample, as is often done? To solve this problem, researchers have turned to “phased” sequencing in order to more accurately represent complex genomes.

In phased sequencing, a researcher tries to capture information about the variation between separate homologous chromosomes by using techniques that enable the isolation and sequencing of each pair of chromosomes. Phased sequencing data sets also enable scientists to distinguish between maternally and paternally inherited alleles, an important consideration when tracing the origins of a genetic disease or condition. Current sequencing

NEW VERSION

SEQUENCHER® 5.3

with Sequencher® Connections

POWER WITH SIMPLICITY

- New RNA-Seq analysis using Cufflinks Suite
- Perform multiple analyses on DNA sequences
- New visualizations

Gene Codes Corporation
T C A G E N E
A G T C O D E S

734.769.7249
GeneCodes.com

technologies and bioinformatics tools do enable short distance phasing—the real problem (similar to sequencing centromeres) is long-range phasing.

One approach to tackling long-range phased sequencing is to develop methods and techniques that actually sort for single copies of chromosomes that can then be labeled or sequenced separately. Wing Hung Wong from Stanford University is interested in phased genome analysis and its applications to medical questions. In 2011, Wong and his colleagues took advantage of flow cytometry and chromosome amplification to sort chromosomes, and then used fluorescent markers to tag DNA from specific chromosomes enabling their phased analysis (4). That same year, one of Wong's Stanford colleagues, bioengineer Stephen Quake, whose lab is interested in single cell analyses, relied on a microfluidic-based approach to initially sort chromosomes for subsequent SNP genotyping and sequence analysis to obtain phased genome data from a single cell (5).

More recently, Karyn Steinberg, along with colleagues including Deanna Church, Richard Wilson, and Evan Eichler, generated a single haplotype genome assembly from a hydatid-

iform mole (6). A hydatidiform mole is unique as it is an abnormal product of conception wherein there is a very early fetal demise and an overgrowth of placental tissue. The result is a DNA source that contains a single

“We are now at the stage where we need to think about the notion of a graph-based reference genome,” remarks Church. Church actually prefers to use the word “model” since she is confident that in the future the way we represent reference genomes will be drastically different from what we do now.

The idea of moving beyond the traditional linear representation of a genome to a graph-based version does have a growing number of supporters—including a group called the Global Alliance for Genomics and

Health. The Global Alliance, lead by David Haussler, a professor of biomechanical engineering and the director of the Genomics Institute at the University of California, Santa Cruz, is a collection of more than 200 institutions working on ways to securely share and present genomic and clinical information. At the most recent American Society of Human Genetics meeting this fall in San Diego, California, the Global Alliance hosted a special workshop aimed at discussing data analysis tools and annotation efforts within the genomics community.

Haussler, along with colleague Benedict Paten, is also a co-investigator on a recently announced \$1 million grant from the Simons Foundation to build a comprehensive map of human variation, which they are calling the Human Genome Variation Map.

“One exemplary human genome cannot represent humanity as a whole, and the scientific community has not been able to agree on a single precise method to refer to and represent human genome variants. There is a great deal we still don't know about human genetic variation because of these problems,” Haussler said in a press release following the announcement of the grant. For their part, Paten and Haussler will take advantage of 300

Mathematical graphs, such as the one above showing variation in the human major histocompatibility complex, could be the future for genome maps. Credit: UC Santa Cruz Genomics Institute.

human haplotype containing primarily paternally derived chromosomes. By sequencing the genomic DNA from a haploid hydatidiform mole, the researchers were able to overcome the issue of allelic diversity faced in diploid genomes without having to pre-sort chromosomes or rely on bioinformatics to obtain phased data.

As phased sequencing efforts continue to improve and as more genetic variation data is being collected by international consortiums such as the 1000 Genomes Project, the question of how best to represent all of this new sequence and genetic variation data on a single “reference” genome becomes more difficult.

Oblio Scientific Consulting, LLC

Specializing in creative solutions to unsolvable problems

Biology, Chemistry, & Education

OblioScientificConsulting.com or oblio_sc@outlook.com

IMPROVING UPON TARGET CAPTURE

Bioo Scientific's NEXTflex™ Pre- & Post- Capture Combo Kit is a library preparation solution tailored for target capture with the Agilent SureSelect system, containing everything needed for library prep, hybridization, and post capture amplification.

Improves Performance of Agilent SureSelect Target Capture Systems

RAPID PROTOCOL

Saves an hour of library prep time

HIGHER % OF ON-TARGET READS

With index-specific blockers

MULTIPLEXING

Up to 96 unique barcodes offering increased flexibility

BETTER COVERAGE

Post-hybridization amplification is performed "off-bead"

Learn more at BiooNGS.com

THE NGS EXPERTS™

Illumina's MiSeq system is an example of a high-throughput, short-read DNA sequencing platform.

complete human genome sequences that have been amassed by researchers from the Broad Institute in Cambridge, Massachusetts to create a single genome representation as a mathematical graph where new genomes can be merged onto a reference genome at points where they match the primary sequence, and genetic variations can appear as alternate pathways. Early discussions on the best path forward for the Human Genome Variation Map are already underway with multiple researchers (including the Global Alliance) contributing ideas.

The speed and depth at which researchers can now sequence DNA is leading to a unique moment in life science research where the genomics community has to step back and figure out how best to share that information with the world. Sequencing difficult regions and finding the best way to represent the data on genome maps will likely usher in a new look for genomics and hopefully reveal many novel insights into the human genome and how genetic variation has shaped the human population.

2. Chin CS, Alexander DH, Marks P, et al. 2014. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*. 10:563-5699.
3. Miga, K.H., Newton, Y., Jain, M., et al. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res*. 24:697-707.
4. Yang, H. Chen X., and Wong W.H. 2011. Completely phased genome sequencing through chromosome sorting. *Proc. Natl. Acad. Sci. USA* 108: 12-17.
5. Fan, H.C., Wang J., Potanina A. and Quake S.R. 2011. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol*. 29:51-57.
6. Steinberg, K.M., Schneider, V.A., Graves-Lindsay, T.A., et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res*. 24:2066-2076.

Written by Nathan Blow, Ph.D.

BioTechniques 58:52-58 (February 2015) doi: 10.2144/000114252

References

1. Koren S., Schatz M.C., Walenz B.P., et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol*. 30:693-700.