October 2012

iCommunity NEWSLETTER

Sequencing the DNA That Encodes Those Tiny Fingers and Toes

Illumina sequencers and a novel approach to haplotyping enable whole-genome sequencing of a human fetus.

Introduction

Since its introduction in the 1980s, non-invasive ultrasound technology has developed rapidly, replacing grainy images with high-resolution video of every heartbeat, yawn, and hiccup. In contrast, whole-genome sequencing of fetal DNA is truly in its infancy. The hurdles are numerous, involving the identification of a non-invasive technique to access fetal DNA and the ability to sequence the fetal genome with an accuracy that meets the stringent requirements for a clinical assay.

Invasive methods to obtain fetal DNA samples from amniotic fluid (amniocentesis) and placental tissue (chorionic villus sampling) have been used for decades, with karyotyping, and more recently, targeted molecular assays able to identify some common mutations. Over the last few years, researchers have refined a non-invasive method that captures cell-free fetal DNA floating in maternal plasma. Combined with next-generation sequencing (NGS), this method is the foundation of several prenatal screening tests to detect major abnormalities in the fetal genome, such as trisomies. However, it doesn't enable the differentiation of fetal DNA from maternal DNA at the high resolution needed to accurately identify inherited alleles or *de novo* mutations, limiting its effectiveness in whole fetal genome sequencing.

The solution may be a sequencing method that leverages haplotypes, groups of variants that reside on the same chromosome. Haplotype information is used in a range of applications, including forensics, population genetics, and *de novo* genome assembly, to interpret genomes and determine genetic diversity and ancestry. A team led by Jay Shendure, M.D., Ph.D., Associate Professor of Genome Sciences at the University of Washington, and doctoral candidate Jacob Kitzman, developed a haplotype-resolved genome sequencing method that combines the throughput of next-generation sequencing with the contiguity information provided by large-insert cloning[1]. Recognizing its potential value in prenatal genetics, the team used the method and the HiSeq® 2000 system to determine the whole-genome sequences of two fetuses, one at 18 weeks and the other at 8 weeks gestation. The method successfully predicted fetal inheritance at 2.8×10^6 parental heterozygous sites with 98.1% accuracy[2].

Jacob Kitzman is a doctoral candidate in the Department of Genome Sciences at the University of Washington.

iCommunity spoke with Mr. Kitzman to learn more about the method and the improvements that will be necessary before it can be used in a clinical setting.

Q: How are you using NGS in your laboratory?

Jacob Kitzman (JK): We're using NGS where it's traditionally shown its value, such as discovering genetic variation associated with a disease state. We're also involved in technology development, where we're looking at types of analyses that NGS currently can't do well such as sequencing haplotypes and figuring out ways to improve them. Standard NGS approaches fragment DNA into very small pieces and mix them together, making it difficult to put those pieces back together again. This comes up in the context of *de novo* genome assembly, as well as in separating the two haplotypes present in every human genome.

Q: What's the value of haplotype information?

JK: Haplotype information is very useful in areas ranging from genetic ancestry and population genetics studies, to cancer genetics. Past approaches have relied upon statistical association between blocks of SNPs, but these are not effective for rare or *de novo* alleles and may lack the accuracy required for clinical sequencing projects.

In this study, we sought to predict the complete sequence of the fetal genome from circulating cell-free DNA in maternal plasma. We needed to know the mother's haplotype to determine which SNPs were from the fetus and which from the mother.

Q. Why did you choose prenatal non-invasive genome sequencing as the first application of your haplotype-resolved genome sequencing method?

JK: We knew there was potential use for this method in a number of applications, but the introduction of several sequencing-based prenatal screening applications motivated us to look closer at prenatal genetics. Those tests only detected disorders caused by large abnormalities, such as trisomies. We felt it would be equally valuable to develop a method to comprehensively detect the whole host of disorders caused by point mutations. Although these disorders are individually rare, they combine to affect a large number of births.

Q: You used cell-free DNA in maternal plasma as a non-invasive method for obtaining fetal DNA. What is cell-free DNA?

JK: We all have cell-free DNA floating in our bloodstream, whether it's naked or part of cellular debris. In a pregnant woman's blood, fetal DNA makes up about 10% of the total circulating cell-free DNA. That makes it difficult to differentiate maternal and fetal DNA, and to detect fetal alleles that are at low frequency among the primarily maternal mixture.

Knowing the maternal haplotype enables us to detect the fetal DNA against the very high background of maternal DNA. Not only is the percentage of fetal DNA in maternal plasma small, but there are also multiple copies of the same gene: two copies from the maternal genome (one from her father and one from her mother), and the one copy the mother passed to the fetal genome. Looking at each site individually doesn't provide a very robust statistical signal for which copy of the gene she passed on. Knowing the maternal haplotype enables you to accumulate SNP evidence along the haplotype blocks, rather than looking point-by-point along her DNA, providing a more confident signature of which gene copies she passed down to the fetus.

Q: Describe the haplotype-resolved genome sequencing method.

JK: Conceptually, the approach takes large chunks of DNA and carries them through the sample preparation process without fragmenting them so the haplotype structure is maintained and variants remain physically linked together prior to sequencing. The specific technique we use is called fosmid cloning, where the library is cloned into a fosmid vector which acts as a substrate for DNA sequencing.

I'll use this study as an example. To determine the maternal haplotype, we first sequenced maternal DNA from blood to 32× coverage. To create haploid subsets, we randomly sheared the maternal DNA and constructed a single, complex fosmid library of approximately 40 clones. We then split a portion of the library into 192 pools, each providing ~5% physical coverage of the genome. Using the Nextera® library prep kit and the HiSeq system, we directly phased 91.4% of 1.9 × 106 heterozygous SNPs into long haplotype blocks, preserving their long-range continuity.

Q: Did you determine the father's haplotype as well?

JK: We would have liked to, but we only had access to the fathers' saliva samples for this study. DNA integrity is crucial for haplotype result sequencing, with blood yielding more intact DNA than saliva. We sequenced the paternal DNA in the saliva sample to 39× coverage, identifying 1.8 × 106 heterozygous SNPs.

Q: How did you classify the maternal and paternal variants?

JK: We sequenced the maternal and paternal DNA libraries on the HiSeq at high coverage, enabling us to see common variants as well as variants that are rare within the population or private to each parent. We could then separate the variants into several classes. The first class is where the mother is homozygous for a given allele and the father is homozygous for a different allele. Those cases are easy, because the fetus will be an obligate heterozygote for that site. The trickier variant sites are the ones where the mother is heterozygous and the father homozygous for the same allele or vice versa. Less interesting from a disease standpoint are variant sites where both parents are heterozygous, because those tend to be common alleles. We compiled a list of variants for all these categories.

Q: How did you identify which variants had been passed down to the fetus?

JK: First, we sequenced the cell-free DNA in the maternal plasma, which is a mixture of fetal and maternal DNA. We built a high complexity library from only about 5 ng of DNA and sequenced it on the HiSeq to a unique sequence coverage of about 80×. We then looked at the sites where there was a maternal or paternal variant that might have been passed down. If we saw alleles in the maternal plasma sequences that matched paternal variants, but that we knew the mother didn't carry, then we knew that the fetus inherited those paternal-specific alleles and we could infer the fetal genotype at those positions. We had enough coverage that if we didn't see the paternal allele in the maternal plasma sequence, we could infer with over 95% accuracy that the father didn't pass down that allele.

We believe the accuracy could be improved simply by sequencing the maternal plasma even more deeply in the future. Basically, we're using the sequencer as an adding machine. If we had sequenced the cell-free DNA to 160× and we still didn't see the paternal-specific allele, then we would be even more confident that the father hadn't passed down that allele.

> "The HiSeq provided us with high coverage and accurate data for the maternal and paternal genomes, and the fetal and maternal DNA in the maternal plasma samples."

On the flip side, in cases where the mother is heterozygous and the father is homozygous for the same allele, how do we infer the maternal alleles passed down to the fetus? That's more challenging and represents one of the values of having haplotype information for this application. We looked across the length of the haplotype in the maternal genome and counted whether on aggregate, one of the maternal haplotypes is over-represented among the plasma sequences versus another one. The haplotype present in excess of its expected abundance is then predicted to have been passed to the fetus.

Q: How did your predicted fetal genome sequences compare with each child's genome?

JK: This was a retrospective study and we were able to sequence the offspring using cord blood samples in order to validate our predictions. We looked at every individual site where one of the parents might have passed down a variant—not counting the sites where the parents had the same genotype—and among the sites coming from mother we obtained 99% accuracy. We think with future improvements that could be pushed above 99.7%, if not higher.

Q: You also looked for de novo *mutations in the fetal DNA. How did you identify them?*

JK: It can be a challenge to find *de novo* mutations even if you have a mother:father:fetus trio that you've whole-genome sequenced at high coverage; here the fetal genome is at very low coverage amongst a background of maternal DNA, so finding *de novo* sites represented quite a challenge. We started with more than 20 million candidate *de novo* sites and by applying simple thresholding, such as requiring multiple reads, high base qualities, filtered for repeats, etc. we ended up with about 4,000 *de novo* calls, fewer than 40 in protein-coding genes. By improving the filtering and increasing coverage for higher sensitivity, we'll be able to more easily and accurately winnow out the true *de novo* mutations in the future.

Q: Did you sequence the fetal DNA of the 18-week gestation sample first, before analyzing the 8.5-week sample?

JK: We decided to sequence the samples simultaneously, mainly because we wanted to show the method was effective for both. We also weren't certain what the percentage of fetal DNA among the maternal plasma was going to be, especially for the 8.5-week sample.

Q: Is it important to verify the percentage of fetal DNA before the sequencing run?

JK: Yes, we want to make sure the quality of the fetal DNA libraries is good before we invest the time and expense of performing whole-genome sequencing at greater than 80-fold coverage. Since the study, we've been using the MiSeq to quickly sequence libraries that we've built from the maternal plasma samples, and have obtained very good estimates of the fetal content percentage. This is similar to using fetal-specific markers, such as Y chromosome sequences in the case of male fetuses, but it works for fetuses of either sex. It also assists with quality control measures such as determining library complexity.

Q: What's been your impression of the MiSeq system?

JK: It's been a very positive experience so far. The MiSeq turnaround time is excellent and the throughput is high. It gives us confidence in whether our library is going to be good or not, saving us money and helping prioritize samples for the HiSeq. It's also easy to use, enabling grad students and technicians to run it without assistance. It's really great to have in our lab.

Q: What surprised you about this research?

JK: We were surprised by how valuable haplotype information was and the difference in accuracy of inferring the fetal genotypes with haplotype information versus without it. That was especially true for the 8.5-week gestation maternal plasma sample where the overall sequencing depth was substantially lower than for the 18-week gestation sample, 50× versus 80× respectively. The percentage fetal content was also quite a bit lower, so on average most places in the 8.5-week genome we only had one read from the fetal genome against a large background of maternal DNA. Despite that, we were still able to achieve over 90% accuracy in inferring the genome of the younger fetus.

Q: Would there have been any benefit to having paternal whole-blood samples for haplotyping?

JK: Absolutely and that's something that we're looking at for a follow-up study. Having paternal haplotypes, instead of viewing genomes on a variant-by-variant basis, will enable us to look across long blocks of the paternal genomes and increase the accuracy quite a bit.

Q: What's the difference between looking at haplotypes versus alleles?

JK: Haplotypes provide much higher data resolution, permitting more accurate inference of maternally transmitted alleles. The deeper we sequenced, the clearer the separation was between haplotype blocks. As the throughput on the instruments continues to go up, this method will become more and more accurate.

> "Since the study, we've been using the MiSeq to quickly sequence libraries that we've built from the maternal plasma samples, and have obtained very good estimates of the fetal content percentage.

Q: How long did it take you to perform whole-genome fetal DNA sequencing using your haplotype method?

JK: If you followed the same steps we performed in this study, you could complete them in three weeks, but that would be pretty optimistic. There's a lot of room for improvement in the experimental and analytical steps that could reduce the workflow to less than a week or two. We're pursuing those improvements and so are a lot of other people.

Q: How did the HiSeq enable you to perform this study?

JK: It provided us with high coverage and accurate data for the maternal and paternal genomes, and the fetal and maternal DNA in the maternal plasma samples. The higher coverage meant there was a cleaner separation between two haplotypes, making it easier to see those variants that had been passed down to the fetus.

There are also a variety of different third-party analysis tools for the HiSeq that were very useful, such as different aligners, packages for calling variants, and data visualization software.

Q: What do you think needs to happen for this method to be relevant in a clinical setting?

JK: There are a lot of improvements that need to be made so this method can be performed outside of a research lab. We'll need to simplify the steps, speed up the workflow, and improve the methodology to increase accuracy in order to make the process more suitable for a clinical laboratory. Clinical laboratories will face additional challenges about how to interpret and report variation.

Q: What is the challenge in interpreting these results?

JK: When you sequence an individual's genome, you get a list of variants and it's difficult to say with certainty whether any of them will predispose that person to disease. Some mutations substantially increase the chances of having a disorder, while others might have no effect. It's difficult enough to figure that out in a research setting. The framework just isn't in place to routinely report variant results in the clinic. This is a challenge faced by clinical sequencing applications in general, not just in the prenatal diagnostic space.

I think this method will initially be most useful for looking at panels of Mendelian disorders where we know the underlying gene. Then, if we see a mutation in this gene that's linked to a specific disorder, the result is more definitive.

Q: How can the haplotyping aspect of the method be made easier so it can be performed in the clinical laboratory?

JK: Fosmid cloning is a well-established technique but it can be a bit cumbersome and time consuming. We're trying to identify ways to do something similar using in vitro techniques or other methods that could be automated.

Q: What's the next step in your research?

JK: We were really pleased with the level of accuracy that we obtained using the method. The fact that we were able to get up to 99.7% accuracy just by looking in large haplotype blocks is promising. We're now focused on process and methodology improvements, finding ways to do this quickly without the cloning steps that we used in the paper and trying to increase the coverage and refine the analysis to improve the accuracy.

Q: Will you continue to use the HiSeq and MiSeq systems?

JK: Yes, HiSeq will remain the work horse for everything and we'll use the MiSeq for quality control preview snapshots of libraries before putting them on the HiSeq.

References

1. Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan PH, et al. (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29: 59–63.
2. Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, et al. (2012) Noninvasive whole-genome sequencing of a human fetus. Sci Transl Med 4: 137ra76.

FOR RESEARCH USE ONLY

Pub. No. 070-2012-017 Current as of 05 October 2012