Good morning everyone! Thank you for being here. This is the public portion of my PhD defense for dissertation titled:

My talk is broken into three sections. I will start with a brief primer on human genetics, then I will summarize the results from my dissertation. I developed an approach for identifying copy number variation in exome sequencing, and will summarize the results from that work. Then, I will talk about detecting fetal genotypes from cell-free DNA in maternal circulation.

My talk today revolves around human genetics. Our hereditary material is DNA. In humans, our DNA is organized into long contiguous strands called chromosomes. In normal circumstance, humans have 46 chromosomes, shown here, 23 we inherit from our biological mother and 23 we inherit from our biological father.

Our DNA works by coding for proteins, which control cellular functions, and ultimately our phenotype. Primary protein structure is defined by a linear chain of amino acids that fold into three-dimensional structure. And that three-dimensional structure is what defines the function of the protein.

DNA codes for protein through transcription into RNA, then translation of three base-pair codes called codons into amino acids. Therefore, changes to the DNA sequence can disrupt protein structure, causing phenotypic changes and disease.

For the purpose of this talk, I need to highlight one additional complexity. In humans, only a portion of the gene sequence codes for protein, and the protein-coding segments called exons are separated by segments called introns. This requires RNA processing to remove the intron segments and stitch together only the protein coding exons. In the image here, we see the gene transcribed, then the dark purple introns are spliced out giving a final messenger RNA with only exons to be translated into protein.

The first part of my work revolves around a specific type of DNA variation, where people have more or less than the 2 copies of DNA we typically inherit.

Shown here is a cartoon of gene. On top we see the gene has 3 exons, which are the protein coding portion, and the individual has two copies of each exon as we would expect. However, we know DNA damage and errors in DNA replication can lead to duplications and deletions. The second individual represented in the middle has a heterozygous duplication of the second exon. The third individual has a homozygous deletion of the third exon. You can imagine resulting protein from this chromosome here with a duplication of exon 2, would have a much different structure and function than the normal protein. This heterozygous example may or may not cause phenotypic changes, because the individual has one copy of the correct gene sequence. Below, in the homozygous example, both copies of the gene are disrupted by the deleted exon.

In a study looking a 140K individuals covering 384 genes, Truty et al. found almost 3000 examples of copy number variants. And while copy number variants were rare, they represented almost 10% of clinically significant findings.

If we look closer at their results, over half of the observed copy number variation was smaller than a gene, with many variants covering only a single exon. This work demonstrates small copy number variants exist, and can have clinical implications, but we know very little about the genome-wide prevalence of exon-level copy number variants. My work focuses on how we can detect exon-level copy number variants in exome sequencing – expanding from hundreds of genes to thousands.

So, what is exome sequencing and why do we do it? Exome sequencing uses engineered DNA fragments called baits to identify and isolate DNA containing specific sequences. We do this for efficiency -- the exome represents only 1-2% of our DNA – and for ease of interpretation. We know disrupting the protein code can change phenotype – it is much more difficult to assess DNA changes outside of the exome. The problem, however, is not all of these baits perform the same.

You can see here the distribution of sequencing depth across an exome we sequenced – not all exons get the same number of sequencing reads. Therefore, we cannot tell if someone has more or less than 2 copies of an exon by simply comparing the sequencing depth to neighboring exons. Instead, we typically compare across individuals. As a toy example, say we look at sequencing depth for an exon across 10 people. One person has 50 sequencing reads where the other 9 average 100. We could make a good guess the person with only 50 reads has a heterozygous deletion of that exon.

Many software packages exist for detecting copy number variation in exome sequencing. I focused on ExomeDepth because it’s the most widely used. ExomeDepth identifies similar individuals from a large pool of exomes, then makes an initial estimation of copy number using a beta-binomial model. The initial estimates are fed into a Markov chain to refine the estimates and identify the boundaries of the variant. Therefore, the algorithm relies on the added power of analyzing contiguous exons, and is not expected to perform well for single-exon variants. But, before asking if we can improve on the algorithm side, I wanted to ask if we can better capture the exome for copy number analysis.

We know most of the variance in sample-to-sample read depth comes from small changes in capture efficiency between reactions. Therefore, we asked the question, what happens to the sample variance when we capture samples simultaneously? The typical workflow is shown above in navy with each line representing a sample. We start by barcoding samples, then we perform exome capture on each sample before they are pooled for sequencing. Alternately, we can pool the samples after barcoding and capture the pooled samples in a single reaction. I will refer to the typical workflow as independent capture and the alternate workflow as multiplexed capture.

To look at the impact of multiplexed capture on inter-sample variance, we collected exomes using multiplexed capture and compared those results to exomes from the NCGENES project. We also used two different capture platforms, one from Agilent and one from IDT. The table gives the pool name, the capture platform, the number of samples in the pool, the median reads per exon across the pool, the median total reads across the pool, and the relative standard deviation of the total sequencing depth. The crosses indicate samples within the pool were captured independently.

This figure gives the results from modeling inter-sample variance by fitting the Dirichlet distribution. Each line represents a pool. The x-axis shows the range of sequencing depths across the pool, with the point indicating the median. The y-axis shows the fitted alpha 0 parameter, with higher alpha 0 indicating lower variance. To compare multiplexed to independent capture, I randomly selected size-matched pools of independent exomes indicated in navy. IC indicates independently-captured pools, MC indicates multiplexed capture. With one exception, we see much less variance in the multiplexed pools. However, you can appreciate the multiplexed pool where we did not see any reduction in variance had a very poor library balance.

If we return to the table, and focus on the relative standard deviation, we see the multiplexed pool with poor variance had over double the RSD of any other pool. We repeated capture and sequencing on the same 16 samples, the IDT-RR pool here, and found when we have an appropriately balanced pool we see the same decrease in variance.

To examine the inter-sample variance another way, we can directly measure the mean and variance across a pool at each exon. This plot shows the mean along the x-axis and the variance along the y-axis for the Agilent capture system. Carolina blue shows the multiplexed pools and the navy shows the same randomly-selected independent pools. In a perfect Poisson process, we expect the mean to equal the variance indicated with the dashed gray line. We see the multiplexed pools show a more Poisson-like process, indicating less inter-sample variance.

We see the same results for the IDT pools. And the multiplex pool overlapping the independent pool is, again, the pool with poor library balance. How does the decreased variance impact ExomeDepth?

Recall ExomeDepth requires well-matched controls for analysis. The software authors suggest reliable results require at least 10 highly-correlated control samples. I wanted to ask if multiplexing the capture is sufficient for defining appropriate controls? On the y axis I have plotted the number of controls selected for each sample in the pool. Independent pools are plotted as triangles, multiplexed pools are plotted as circles. For most of the multiplexed samples we found appropriate controls sets within the pool. For independent captures, we see most samples found sufficient controls for the NCGENES pool, but not for the IDT-IC pool. This highlights the need for a large pool of controls to select from when using independent capture.

I highlight the difference in pool size, here. The y-axis gives the proportion of available samples selected as a control. Again, ExomeDepth found sufficient controls for most NCGENES samples, but that required over 100 exomes, whereas the multiplex pools often used every sample in the control set.

Finally, we can look at the dispersion modeled by ExomeDepth. Dispersion is the amount of variance above the expected Poisson variance. We find often multiplexed pools show lower dispersion, even after selecting appropriate controls – suggesting the control selection does not fully compensate for independent captures.

So, can we better capture exomes? I think we can. Based on my results, I think all exome captures should be multiplexed over 12 to 16 samples. I think broad implementation of multiplexed capture with adequate library balance and sequencing depth will dramatically increase our knowledge about exon-level copy number variation.

Next, I asked if we could improve on the algorithm. Based off RNA sequencing methods, I developed the mcCNV algorithm. mcCNV assumes multiplexed capture. It is an iterative algorithm that uses a shrinkage estimator for dispersion and estimates copy number using the negative binomial distribution.

I will start by showing the comparison to ExomeDepth on simulated data. On all of these plots, the x axis shows the ExomeDepth performance and the y-axis shows mcCNV. Each numbered point gives a sequencing depth in millions of reads. The left panel shows ExomeDepth using default settings, and the right panel shows ExomeDepth using settings matching the simulation parameters. This first set of plots shows the Matthew’s correlation coefficient, which is analogous to area under the ROC curve. On the left, we see mcCNV performs uniformly better when we do not know the true prior for having a copy number variant. Performance is similar, but slightly better for ExomeDepth when given the true prior probability.

Here are the same plots showing the sensitivity.

And here are the same plots showing the false discovery rate. In both instances, mcCNV controls the FDR much better than ExomeDepth.

For one of the multiplexed pools, we also performed genome sequencing with 50x coverage. I then used the best practice pipeline defined by Trost et al. to estimate copy number variants from the genome sequencing results.

Here is the summary of CNV calls from the three approaches. In rows we have individual samples and the number of calls per sample by algorithm in total and broken out by duplications and deletions. MC indicates mcCNV, ED indicates ExomeDepth, and WG indicates the genome results. Almost uniformly, mcCNV predicts the fewest variants. Two notable exceptions are samples 790 and 851, where both mcCNV and ExomeDepth predicted thousands of deletions not seen in the genome results.

Here, I am showing Matthew’s correlation coefficient, sensitivity, false discovery rate, and positive predictive value using either mcCNV or ExomeDepth to predict the genome calls. I show the values for all calls, and broken out by duplications and deletions. Additionally, I show the values including and excluding the two samples with disproportionately high deletions. We see almost uniformly better performance with mcCNV. Both exome methods seem to under-call duplications. Alternately, the genome approach may over-call duplications. Overall, the performance was not as great as the simulations suggested.

These venn diagrams show the call overlap for the three approaches for duplications on the left and deletions on the right. I am not including the two samples with excessive deletion calls.

So, can we better analyze exomes? I would say possibly. I think the fairest thing to say is mcCNV performs very comparably to ExomeDepth. However, mcCNV does not require any prior information. Since we have essentially no genome wide data on the prevalence of small exon-level copy number variants, I think this is a big positive.

I am going to totally shift gears now and talk about my second project looking at noninvasive fetal sequencing.

In 1997, Dennis Lo and colleagues discovered fetal and placental cell-free DNA in maternal circulation during pregnancy. This discovery of cell-free fetal DNA revolutionized the field of prenatal genetics, and now women can get noninvasive blood testing many genetic conditions. Most commonly, women get testing for trisomy’s 13, 18, and 21, but more recently testing for dominant single gene disorders was released by Baylor Genetics and now the NHS in the UK is offering testing for cystic fibrosis noninvasively.

Despite the numerous advancements in noninvasive fetal sequencing, no one has been able to demonstrate reliable fetal genotyping without additional sequencing information. The noninvasive cystic fibrosis testing in the UK requires sequencing both the mother and the father, in addition to the cell-free DNA. Ultimately, we would like to demonstrate fetal genotyping purely from cell-free DNA. Doing so would bypass many ethical issues and enable large-scale testing.

Performing fetal genotyping from cell-free DNA requires two estimation steps. In standard genotyping, we simply observe the proportion of minor alleles. In this example, we would observe about half of the reads contain T and half contain A, suggesting the individual is heterozygous at this locus. However, cell-free DNA is more complicated because we are observing DNA from two sources. The majority of cell-free DNA in maternal circulation is derived from maternal cells, with only a small fraction coming from the fetus. So, before we estimate genotypes, we must first estimate how much of the total cell-free DNA is fetal-derived, which we call the fetal fraction. Then, using the estimate for the fetal fraction, and the observed proportion of minor alleles we can calculate probabilities for each possible pair of maternal and fetal genotypes. In this illustration, it is mostly likely the mother is heterozygous and the fetus is homozygous for the A allele. However, that does not preclude the possibility the fetus is actually homozygous for the T allele.

The expected proportion of minor alleles, T to A in the previous example, is well-defined given the fetal fraction. In other words, each pair of genotypes has a distinct expected fraction of minor alleles when the fetal fraction is known.

Here, I show expected proportion of minor alleles along the y-axis for different fetal fractions along the x-axis. Notice the light blue and light orange bands are well-separated from the rest. These, sites, where the fetus is uniquely heterozygous can be exploited to estimate the fetal fraction.

I wrote an algorithm for performing noninvasive fetal genotyping using these principles. In the first step, I perform an empirical Bayes routine to identify sites with unique fetal heterozygosity. I then calculate the observed fetal fraction at each of these sites and use the median value to estimate the overall fetal fraction. With the fetal fraction estimate I build a straight-forward maximal likelihood model using the binomial distribution.

I used my algorithm to analyze noninvasive exome sequencing for three pregnant women. These cases were recruited by genetic counselors at UNC as good candidates for genetic testing prior to any genetic diagnosis being made. Case 1 had five prior pregnancies affected with X-linked Menke’s syndrome. She was later found to have a single exon deletion in the ATP7A gene. Case 2 had sonogram findings suggestive of osteogenesis imperfecta, and was eventually found to carry a pathogenic variant in the P3H1 gene. In case 3, the fetus was found to have clubbed feet and upper limb arthrogryposis bilaterally. We have still not found any genetic cause for the fetal presentation.

Here are the results for Case 1. The histograms show the distribution of observed allele proportions. The overlaid lines show the theoretical distributions. We see a large degree of overlap, particularly in the center, and that the observed distributions poorly match the expected distributions.

The same is true for Case 2. We correctly genotyped the fetus as homozygous for the pathogenic allele, but as you can see, that was purely by chance. The overlapping distributions here preclude any reliable interpretation of these results.

Finally, here is Case 3, showing an even greater proportion of the results falling outside the expected distributions.

For Case 3, we also have individual exome sequencing on the mother, father, and newborn. With the individual sequencing, I looked at how accurately I could predict fetal genotypes. Down the rows I show the true fetal genotype, with the estimates from the cell-free DNA across the columns. Overall, we found a very poor 50% genotyping accuracy.

In the end, these results are not surprising. The median sequencing depth in our three cases ranged from 150 to 330. Here I am showing the 95% confidence intervals for binomial proportions representing a sequencing depth of 500. We can see a high degree of overlap until fetal fractions surpass 20%. If we think about the testing range being in the late first or early second trimester, even sequencing depths of 500 are insufficient. If we plot the fraction of overlap, here, for different sequencing depths, we get a plot that looks like this.

Along the x-axis I am showing sequencing depth, with the degree of distribution overlap (a proxy for error rate) on the y-axis. The lines here represent fetal fractions increasing from 5% up to 25%. For the low fetal fractions observed within the testing window, we need very deep sequencing for reliable results, likely greater than 5,000x.

When we initially submitted this work, reviewers asked us to integrate fragment length into our estimation. That criticism derived from previous working showing fetal-derived fragments are about 15 bases shorter than maternal fragments, on average. Rabinowitz and colleagues used this finding to develop a genotyping algorithm called Hoobari integrating fragment length.

So, I went back and analyzed fragment lengths. Using the individual sequencing results from Case 3 I identified alleles unique to either the mother or the fetus, then I collected the cell-free reads supporting only those unique alleles. Here I show the distribution of read lengths for the maternal fragments in navy and the fetal fragments in Carolina blue. I also found a slight left shift in the length distribution, but the shift was not absolute and I observed a higher proportion of long fragments derived from the fetus as well. On the right, I am showing the cumulative distribution. One group isolated only fragments less than 150 basepairs for their analyses, but that would throw out 80% of the data and compound the sequencing depth problem we observed.

Finally, I looked at the correlation between the observed allele ratio on the y-axis and the proportion of reads less than 140 basepairs on the x-axis. If incorporating fragment length would provide more robust results, we would need to observe at least a moderate positive relationship between the proportion of short reads and the allele ratio. In our three cases, we observe no relationship at all. Furthermore, the Hoobari algorithm found very little change in genotyping accuracy when incorporating fragment length (and, in fact, found worse performance in some instances).

In summary, noninvasive exome sequencing would require cost-prohibitive sequencing depths (roughly $400K in current sequencing costs), and we find no benefit to correcting for fragment length. I am currently working on a project with Neeta Vora, using what we learned here, to try a more targeted approach in 100 women.