A Tour of Next Generation Sequencing Technology:

Inside the Washington University Genome Center

Scene 1: Introduction. In Elaine Mardis's office.
Libby knocks as Dr. Mardis is sitting at her desk typing on her computer.

E. Mardis: Hi, Libby. Come on in.

Libby: Hi, Dr. Mardis.

EM: It's nice to see you again. A lot has changed since you last visited! Did you notice on your way in that we've changed our name? We're now called the Genome Center.

L: Why the change?

EM: We've entered a new phase in genomics. We have a complete human genome sequence. In the past 5 years sequencing technologies and applications have been changing rapidly. Although we're still sequencing the whole genomes of new organisms, we're also involved in a variety of other projects that use DNA sequencing as a tool to answer a research question.

L: I'd love to hear about that. First, can you tell me what new genomes you've sequenced?

EM: Sure. We've been pretty busy doing a lot of genomes. In the past 5 years, we've sequenced the genomes of the chimpanzee, the platypus, the zebra finch, the Rhesus Macaque, and many bacterial genomes.

In fact, if you take a look at this graph, it shows the exponential increase not only in the number of runs we've been able to do on our sequencing instruments, shown in red, but also the number of reads we've been able to generate, in thousands, shown in blue. In the last two years, in 2007 and 2008, the number of runs per instrument has dropped, but the number of reads has gone up exponentially. And that's why I was interested to have you come in today and hear about next generation sequencing technologies and how we've been able to achieve this dramatic increase in productivity with fewer machine runs.

Our production has increased exponentially. We can now sequence the genome of a new bacterium for one to five thousand dollars, depending on the platform we use. Ten years ago, sequencing the same bacterial genome would have cost hundreds of thousands of dollars. This means we have decreased our production costs over 100 fold in just 10 years.

L: Wow! That's quite a change. How have you increased sequencing production while decreasing costs?

EM: Several new sequencing technologies have been developed, which we call "next generation sequencing." Next generation sequencing technologies are so much faster and cheaper that we can do a whole range of new projects. For example, we can sequence the DNA or RNA from tumor cells to look at the genome and gene expression in certain types of cancer. We've completed a project to compare DNA from two types of cells in a patient with acute myeloid leukemia; we compared DNA from non-cancerous cells with DNA from tumor cells and discovered mutations that were specific to the cancer. Experiments like these will help bring genomics into doctors' offices for diagnosis and treatment.

We can also sequence mixed populations of organisms in what are called "metagenomic" experiments. For example, we can sequence DNA isolated from seawater or from the human gut to see what type of bacteria are growing in each place. These types of experiments would be too expensive to perform on a large scale using the traditional Sanger sequencing methods in the past.

L: So, what are the main differences between the old Sanger technology and the new next generation sequencing technologies?

EM: These technologies are based on two principles: first, massively parallel sequencing, which means a huge number of sequencing reactions are performed and detected at the same time; and second, sequencing by synthesis, which means we monitor the sequencing reaction as it occurs -- there is no electophoresis step.

The new sequencers work directly with the DNA fragments from a genome, and don't require cloning or any bacterial work, which decreases the production pipeline time significantly. In Sanger sequencing, which you saw in your last tour, a genome will be fragmented to create a library of cloned DNA fragments. DNA would be prepared from a selected set of overlapping clones, and sequenced using ABI machines that separated DNA fragments using gel electorphoresis. By contrast, next generation techniques simply amplify the genomic fragments using PCR and then determine the sequence by monitoring the incorporation of nucleotides as DNA is synthesized from the template. Most of the steps in sample preparation are completely automated and performed by machines. The sequencing machines have much longer run times—sometimes as long as 10 days!—so a sequencing lab technician can keep several machines running at once.

Another major difference between the Sanger technology and the next generation sequencers is that the new technologies give us shorter readlengths for each individual sequencing reaction. This means that the analysis is much more computationally intensive than in the past, and as a result, we employ far more personnel trained in bioinformatics than ever before.

L: You said there were several new technologies. What are they?

EM: The next generation technologies we're currently using at the Genome Center include the Illumina GAIIx and 454 Titanium sequencers.

L: What are the differences between them?

EM: Let me first explain the differences in terms of cost and productivity; then you can talk to our outreach director to hear all about the details of the technologies.

If you look at this chart, it compares the ABI 3730 XL capillary sequencer to the Illumina GAIIx and the 454 Titanium sequencers. As you can see, we have only 30 ABI 3730 XL sequencers remaining at the Genome Center, whereas last time you visited there were 138 of these sequencers. By contrast, the number of Illumina and 454 sequencers are much lower. Where the ABI 3730 sequencers can only produce on the order of 60,000 base pairs per run, the 454 sequencer produces 400,000,000 base pairs per run, and the Illumina sequencer over 18,000,000,000 base pairs per run. This is a very dramatic increase in the amount of sequence data produced, and so we need fewer sequencers to produce the sequences for our experiments.

The most astonishing component and contribution of next generation sequencing is illustrated in the last column of the chart, namely the time to complete the 3 billion base pairs of the human genome. Whereas with the 3730 sequencers, over six years would be required to generate this amount of data, the 454 and Illumina sequencers can complete it in less than a week.

	ABI 3730XL	Illumina GA_{IIx}	454 Titanium
Number of machines	30	38	8
Sequencing method	Sanger	Reversible terminator synthesis	Pyrosequencing
Cost of machine	New $350K Now $25K	$540K	$500K
Runtime	1 hour	3-10 days	7-8 hours
Readlength	700 bp	75-100 bp	400 bp
Bases per run	60,000	18,000,000,000	400,000,000
Bases per day	1,300,000	2,500,000,000	800,000,000
Relative cost	100 - 500X	1X	10X
Time to complete 1x human genome	2,307 days (6.3 years)	1.2 days	3.75 days

These numbers, of course, are changing rapidly; readlengths are getting longer and costs are getting lower. But you can tell that the three platforms have major differences in readlength and total production capacity. The different technology that the machines use for sequencing means that they also have differences in base quality and different problem areas, or biases as we refer to them. All of these factors, as well as cost of production, do play a role in deciding which experiment/platform to use. Depending upon what project we're working on, we use the machines for different applications, and both of the new technologies, as well as the old pipeline, play an important roles in our current production pipeline.

To really understand the details of the new technologies, let me ask our outreach director, Cherilynn Shadding, to show you the machines and to tell you more about how they work.

L: Thanks, Dr. Mardis.

Scene 2: Illumina. In a lab.

C. Shadding: Dr. Mardis has given you a good idea about the differences in production, now I'll tell you about the differences in technology and you can see the machines. Both technologies have some similarities. At the beginning, we start with input DNA. As Dr. Mardis said, that DNA can come from a variety of sources and can be from a single organism or from a mixture of organisms. Before starting, the input DNA is broken up into many short fragments of DNA. The first step of the sequencing process for both machines is to make the input DNA into a library and to amplify each fragment, making many copies of each piece of input DNA. This step doesn't involve bacteria like the old sequencing pipeline, just DNA polymerase. At the end, the sequencing reactions are detected by a camera observing a flash of light. In the middle, there are many differences.

Let's start by talking about the Illumina technology. Illumina technology uses two machines called the Cluster Generator and the Genome Analyzer.

If we have full-length genomic DNA, the first step in making a sequencing library is to break the DNA into smaller fragments. This is done in a nebulizer in library core. After the DNA is in smaller pieces, the ends are repaired to make them blunt-ended. Then we ligate or attach adapters to both ends of the small pieces of input DNA. The input DNA is represented in gray and the adapters are represented in purple and pink. We add adapters so that we know the sequence at the end of the DNA strands, allowing us to design oligonucleotide primers. Remember that polymerases require primers to start copying a sequence.

A flowcell is like a slide with eight sample lanes etched on it and a coverslip positioned on top that's attached to a collection of tubes. Small volumes of liquid can be pumped through the tubes so that the liquid completely covers the surface of the slide. The machine pumps different amounts of different liquids through the flowcell to automate steps of a reaction.

After the adapters are ligated to the ends of the input DNA, the DNA is denatured to make it single stranded and the sequencing library is loaded into the flowcell on the Cluster Generator. It is attached to several sets of tubes that will deliver all of the reagents and enzymes needed for the reaction.

The surface of the flowcell has a lawn of primers, short sequences that match to the adapter sequences, shown here as short strings of pink and purple dots. When the single-stranded DNA is loaded on the machine, shown here in black, the adapters will form hydrogen bonds to primers with the matching sequences, pink adapters binding to pink primers and purple to purple. Polymerase is added and starts to make a copy of the template strand. The copied strand is shown in grey. Because the copied strand starts with a primer attached to the surface of the flowcell, that strand is bound to the surface of the flowcell. Then the temperature can be increased and the original strand is melted off, leaving a single strand bound to the flowcell surface, surrounded by a dense lawn of primers. There are millions of pieces of input DNA spread out over the surface of the flowcell, and each has a unique sequence. However, we can't yet sequence the DNA because there's only one copy, so we go through a process called bridge amplification—which is just another type of PCR—to amplify the DNA so there are many copies sitting in the same place in the flowcell.

The temperature is decreased so that the DNA stuck to the flowcell forms hydrogen bonds with a primer on the surface of the flowcell, shown here in purple, creating a bridge. Then a PCR reaction mixture is run over the flowcell. The temperature is increased and polymerase makes a copy of the input DNA. The growing strand is seen in gray. We call this bridge amplification, because the copying occurs in the arched region between the two adapter sequences. When the polymerase is finished, there is a double stranded arch with one end of each strand attached to the surface of the flowcell.

Then, just like any other PCR reaction, we heat the flowcell and the two strands denature or separate. In this case, one strand is attached to the surface of the flowcell by the pink primer and the other strand is attached to the flowcell by the purple primer. Then we can go through another temperature cycle. We lower the temperature so that each single stranded piece of DNA anneals to a primer on the surface of the flowcell and forms a bridge. The temperature is increased to allow the polymerase to extend a copy of the DNA. The flowcell is heated, and we have four single-stranded copies of the input DNA. All copies are attached to the surface of the flowcell very close to one another.

This can be continued for many cycles of amplification, leading to a million copies of single stranded DNA attached to the surface of the flowcell in a cluster. Because we started the process by attaching many different pieces of input DNA spread out over the surface of the flowcell, we now have many individual clusters of the input DNA.

L: Why don't you do the PCR in tubes? You can do a lot of PCR reactions a microtiter plate.

CS: This is a very different scale. Even in the best microtiter plates available, we could only do 384 PCR reactions at once. In the cluster generator, we can 200 million pieces of input DNA over the surface of the flowcell. Because the DNA is spread out and attached to different positions, a different PCR reaction occurs at each separate position. The amount of reagents we use is greatly decreased, and we end up with millions of clusters of DNA, each cluster having a particular sequence.

Now, one of the strands is cleaved within the adapter sequence and the reaction is heated to separate the strands. Because one strand is not attached to the flowcell, it floats away. Now each cluster contains only single stranded DNA of a given sequence. The sequencing primer is added, it hybridizes to the corresponding adapter sequence, and the clusters are ready for sequencing. The flowcell is removed from the Cluster Generator machine.

For the next step, the flowcell is loaded into the Genome Analyzer. This machine also has reservoirs attached to a set of tubes to deliver the reagents needed during sequencing, such as buffers, DNA polymerase, and fluorescently-labeled terminator nucleotides.

Illumina and Sanger sequencing are alike. The bases are represented by a pentagon for the sugar backbone, a circle with a P for each phosphate group, and a rectangular box for the base, A, T, G, or C. In the sequencing reaction, every base is labeled with a different fluorescently colored dye, represented by a star, and has a 3-prime terminator, represented by a stop sign on the 3-prime carbon, that keeps the reaction from proceeding beyond a single addition.

Unlike Sanger sequencing, all of the nucleotide bases are labeled with fluorescent dye and have terminators. That means the polymerase adds only one labeled nucleotide to each strand of the sample DNA then stops because of the 3-prime terminator. A laser excites the dye, which flashes with the color corresponding to the base. The color of the flash is detected by a camera.

L: That’s a lot like the old sequencing method! But with the old method, the cycle couldn’t continue after you added the dye terminators. How is this different?

CS: The major difference is that after the first cycle, the dye and the terminator are both removed. Then another round of reaction mix can be run over the flow cell and another base can be added and detected by the camera; then the dye is removed along with the terminator block. The camera records the timing, color, and position of each light flash, and so the sequence of the bases can be determined for each cluster.

The flowcell is attached to tubes that deliver the reagents; it remains stationary with a laser pointing at it. The camera is positioned above the flowcell to capture images of flashes from all the clusters on the flowcell.

L: Reversible terminators sound like a really good idea! Are there any problems with this sequencing method?

CS: The removal of the dye and the terminator is difficult and not foolproof. If the dyes are not completely removed, then two colors will be read by the camera in the next cycle. If we fail to remove all of the terminators, then the signal strength will be decreased because fewer dye-labeled nucleotides will be incorporated in the next cycle. Because of these difficulties, the maximum readlength of Illumina is about 100 bases, and there can be a small error rate of incorrect bases in the sequence reads. Both the Genome Center and Illumina are working to increase the readlength, but for many sequencing projects, this readlength is sufficient.

L: What about paired end reads? Is that possible with Illumina?

CS: Yes. There is a machine that can be attached to the Genome Analyzer called a paired end module. This acts like a small cluster generator to regenerate clusters for paired end sequencing.

After sequencing with one primer, the flowcell can be heated to separate the sequenced strand and then washed off. A couple cycles of bridge amplification regenerate the other strand, and the first strand can be cleaved, denatured, and washed away. Then the sequencing can be repeated using a primer with a sequence that's complementary to the other adapter. Because the clusters remain in the same location, the computer can record the information knowing that this same fragment, and we get paired end reads – sequence information that we know comes from the two ends of the same DNA fragment. This is very useful in assembling the sequence information.

In an 8 day run for one flowcell, we observe about 200 million clusters with a readlength of 75-100 bases each. The overall amount of sequence generated can be as high as 20 billion bases, which means we get 6x coverage of the human genome in 8 days. That means Illumina sequencers can obtain 1x coverage of the human genome in 1.2 days.

L: That's pretty amazing.

CS: It is.

Scene 3: 454. Still in lab.
Now let's talk about the other sequencing technology we're using. This is a 454 machine, made by Roche. The first step is the creation of a library from the input DNA. Just like in Illumina sequencing, the DNA to be sequenced is broken into smaller pieces and adapter sequences, represented here in red and green, are attached or ligated to the ends of the input DNA, which is shown in grey. The adapters have tags on the ends that allow us to separate the pieces of input DNA that have adapters on both ends. We then heat the strands to separate them and mix the DNA with primer-coated agarose beads, PCR mixture, and oil.

When the tube is vortexed, the beads are suspended in little drops of water containing PCR reaction mixture in a sea of oil—this is called an emulsion. This is just like salad dressing. We call the little droplets microreactors, and we want every droplet to contain a single piece of input DNA and a single bead.

L: How do you make sure that only one piece of input DNA is in the droplet with a bead?

CS: We use a dilute solution of input DNA with an excess of beads to make sure that most of the microreactors contain only one bead and one piece of DNA.

Then we amplify the input DNA on the beads. This occurs in a PCR machine. When we go through the PCR temperature cycles, amplification occurs in each microreactor. The surface of the bead is coated with primers much like the surface of the Illumina flowcell is coated with primers. The adapter on one end of the input DNA anneals to the primer on the surface of the bead, and the polymerase in solution extends the sequence from the primer to make a strand of DNA that's attached to the bead. When the tube is heated, the template strand is denatured from the bead, that strand is copied within the microreactor, and the cycle continues. In the end, millions of copies of the input DNA are made in each droplet.

L: Why do you use an emulsion instead of regular PCR or bridge amplification like Illumina?

CS: Remember, the input DNA is in many small pieces and all the pieces have a different DNA sequence. Each of the beads has a novel piece of input DNA. By suspending the beads in oil, we create water droplets each with a single bead, the microreactor. A different PCR reaction occurs in each water droplet so we can have millions of different PCR reactions occurring in a single PCR tube. Emulsion PCR allows us to perform many PCR reactions in a small volume of reaction mixture and serves the same purpose as spreading the reactions out over the surface of a flowcell.

At the end of the PCR cycles, each individual bead has a million copies of the input DNA attached to it. We heat to separate the two strands of DNA; then each bead has a million copies of single stranded template DNA ready for sequencing. After we centrifuge the tubes to separate the beads from the oil, we're ready for the next step.

The next step occurs in the 454 machine. First, we take the beads and run them over a picotiter plate.

L: What's a picotiter plate?

CS: Picotiter plates are like much microtiter plates but there are several million smaller wells, created by etching a glass slide. The wells in the picotiter plate are just large enough so that a single bead with template DNA fits into each well. Then we add a second solution that has smaller beads with enzymes attached to the beads. These smaller beads and fit into the wells around the DNA beads. Then we're ready for the sequencing reactions. The picotiter plate is loaded onto the 454 and attached to tubes that pump liquids with the additional needed chemicals over the picotiter plate.

454 sequencing uses a technology called pyrosequencing. In pyrosequencing, we're observing the addition of a base by detecting the release of a pyrophosphate.

First, a primer is added that binds to the adapter sequences on the free 5-prime end of the single stranded template. Unlabeled, normal nucleotide bases are added—note that there is no dye and no terminator block. The unlabeled nucleotide triphosphate bases are run over the picotiter plate one at a time: G, then A, then T, then C. If the next base in the template DNA is C, then the polymerase adds dGTP to the extending strand. Remember, deoxyguanosine triphosphate, or dGTP, has three phosphate groups. When polymerase incorporates dGTP into the growing DNA strand, one phosphate group is used in the phosphodiester bond to make the sugar phosphate backbone of the DNA and the two-phosphate group is released. This two-phosphate group is called a pyrophosphate. Other enzymes on the beads in the picotiter plate react with the pyrophosphates: the blue arrow represents sulfurylase, which reacts with the pyrophosphate to make ATP; the red arrow represents luciferase, which reacts with the ATP and luciferin to produce a signal in the form of a flash of light.

L: Where does the luciferin come from?

CS: The luciferin is stored in the reagent tubes and is pumped over the picotiter plate with the nucleotides during the reaction. Luciferin and luciferase were originally discovered in fireflies.

The flash of light produced by luciferase is detected by a camera that's mounted next to the plate in the 454 machine.

After the flash of light is observed, the excess nucleotides are washed off and the next base is run over the picotiter plate. Each nucleoitide is run one by one (A, T, G, C) with a wash in between. The camera records the position and timing of each flash of light. 200 cycles are carried out with all four bases run one after another during each cycle, resulting in a readlength of 350-400 bases.

The 454 beads use a primer from only one end of each DNA fragment. But you can create paired end DNA libraries that have an adapter sequence attached to two pieces of DNA that are found 2.5 kilobases apart in the genome. Sequencing of these paired end libraries can give you more information to assemble sequences from novel genomes.

L: Does this machine have any problem areas?

CS: Overall, this machine gives very good quality sequence. Base substitutions errors are very rare because you're only running one base over the picotiter plate at a time and the nucleotides don't have any modifications that might interfere with recognition by the polymerase.

However, a problem occurs when there are many of the same bases in a row. If there are 3 adenosines in a row in the template strand, then the polymerase will add three dTTPs all at once when dTTP is present. If there are 4 As in a row, the polymerase will add four dTTPs, and so on. This will release a larger amount of pyrophosphates and lead to a brighter flash of light. If the number of bases in a row is more than 6, though, the brightness of the flash will not be proportional to the number of bases added. So long runs of a single nucleotide can be misrepresented in this sequence data. This is called condensing a homopolymer run.

Overall, there are 3.6 million wells in each picotiter plate, and most wells contain a single bead. The camera has recorded about 400 bases of sequence data from each well. Up to 400 megabases—that's 400 million bases—of sequence data with very few substitution errors can be obtained in an 8 hour run. We do at least two runs a day in each machine, which means we get 800 million bases per machine per day. That's a little over a quarter of a human genome a day, which means it takes about 3.75 days to get 1x coverage of the human genome.

L: 454 sequencing is pretty good, but Illumina generates more sequence data, faster. Why do you use both?

CS: Well, not every project needs as much sequencing power as the Illumina technology can provide, so the 454 machines are good for smaller projects. Also, as I've mentioned, the two technologies have different problem areas, so some projects will work better with one technology or the other. The shorter readlengths of Illumina sequencing technology can mean it's little harder to assemble the data. This means we have to sequence to a greater coverage with Illumina machines—at least 20x coverage is necessary, sometimes as much as 50x coverage. This raises the cost and decreases the speed of finishing a project. 454 give longer readlengths so analysis can be slightly easier. Both next generation technologies are significantly faster and cheaper than the old Sanger sequencing pipeline, and we use Illumina, 454, and Sanger sequencing, separately and in combination, for sequencing projects at the Genome Center.

Now that you've learned about the technology used by the machines, I'm going to take you to the director of the Genome Center, Dr. Rick Wilson, to talk about new applications of next generation sequencing technology.

L: Thanks, Dr. Shadding.

Scene 4: Applications. In Richard Wilson's office.
R. Wilson: Hi, Libby.

Libby: Hi, Dr. Wilson.

RW: What did you think about the new next gen sequencers?

L: The new sequencing technologies are very different than the old pipeline. Are you still sequencing new organisms?

RW: Of course. We're using a combination of Sanger sequencing, Illumina, and 454 technologies to sequence and assemble the genome of novel organisms. For example, we're working on the genomes of corn, the cat, Hoffman's Two-toed Sloth, the alpaca*,* and the mosquito responsible for spreading malaria, *Anopheles gambiae.*

The new technologies allow us to do a whole range of different research projects. Both of the new sequencing technologies are significantly cheaper than the old pipeline.

L: Are the shorter readlengths of the new sequencing technologies a problem when you assemble and analyze the data?

RW: It depends on what you want to do. Shorter readlengths could be a problem in assembling the genome sequence of a completely new species. But for many interesting studies, what we want to do is resequence the species, using a different individual. For example, we have a very high quality human genome reference sequence. Because of this, if we get a sequence read of 75 basepairs from an Illumina machine, we can use the computer to map where that sequence comes from. In many cases, we can map even short sequences to a unique position in the genome. This means that we can generate the sequence of the coding portion of your genome for much less expense than ever before.

We can also use short sequences for other types of experiments. For some projects, we're interested in studying enrichment of parts of the genome in certain cells. These are more like counting experiments than actual sequencing experiments.

For example, we can isolate RNA from a particular tissue or development stage, then make complementary DNA by using reverse transcriptase, and then sequence the DNA. Even with short sequence reads, we're able to map the sequences back to the genome and measure the levels of gene expression.

In another example, instead of isolating all of the genomic DNA from a cell, we can isolate only the genomic DNA that is bound by a particular protein. So if we're interested in a specific transcription factor, we can use antibodies to isolate DNA fragments that are bound to our favorite transcription factor. We then sequence the DNA fragments with Illumina technology and use the short sequence reads to map the DNA back to the genome. Overall, the experiment measures how many reads come from each location in the genome; so it tells us how often the transcription factor protein is found at each place in the genome. This is called a "ChIP" experiment, which stands for Chromatin ImmunoPrecipitation. It is a new and exciting way to study how proteins interact with DNA on a genome-wide scale.

L: That sounds interesting. Can you tell me more about the metagenomics projects that Dr. Mardis mentioned?

RW: In metagenomics, we're studying DNA isolated directly from the environment or from a community of organisms instead of just one specific organism. Dr. George Weinstock, who's an Associate Director here at the Genome Center, has said that "you are crawling with microorganisms," which refers to the fact that we have many microbes living symbiotically with us. These include bacteria, fungi, protazoans, and viruses. For each person, microbial cells outnumber human cells by a factor of ten to one! Most of these microbes are not harmful and some are beneficial—they help you digest your food, produce proteins, and protect against disease-causing microorganisms. Some microbes are harmful and can cause disease.

So at the Genome Center, we're studying several microbiomes. As part of the Human Microbiome Project, we're sequencing and cataloging entire microbial communities from the nasal, oral, skin, gastrointestinal, and urogenital systems. Next generation sequencing technologies allow us to sequence the entire microbial community from an individual. We can compare it to microbial populations from other people or from different environments in the same individual.

And in addition to creating a reference data set of human microbial communities, we are also sequencing microbiomes from people with disease. Washington University researchers are examining the microbiome in people with Crohn's disease, infants with necrotizing enterocolitis, and children with fever of unknown origin. We would like to find out how the microbiome differs in individuals with these diseases, and that could lead to new methods of prevention or treatment.

L: If you're sequencing many new microbial communities, are you going to sequence more human genomes, too?

RW: Yes, we're working on the 1000 Genomes Project. We're using our Illumina sequencers to sequence the genomes of hundreds of individuals. Right now, public databases only contain genetic information from a handful of people, so by adding the genomes of many people from around the world, this will help us to determine levels of variation in the normal population. Many researchers will use these multi-genome databases to look for common and rare gene variants. They can then use that information in gene association studies to determine the genetic basis of diseases.

We're also interested in variation and mutations in cells that have become cancerous. We study this by examining the DNA from two cell types from the same person—DNA from their normal, non-cancerous cells as well as DNA from their tumor cells.

L: Dr. Mardis mentioned this. What have you found so far?

RW: Well, we've already compared cancerous and non-cancerous tissue in two individuals with acute myeloid leukemia, or AML. We were able to find specific mutations in the cancer cells that had not been identified before. This information can be used to identify mutations that are correlated with disease outcomes, giving the physician a new tool to help decide on a course of treatment or design a new course of treatment.

Also, a large-scale project called The Cancer Genome Atlas, or TCGA, has been started. Our first project in that was to sequence and analyze over 206 cases of glioblastoma, a form of adult brain cancer. We discovered that current treatments may actually lead directly to mutations that cause treatment resistance, and were able to suggest a new avenue of treatment that may prevent treatment resistance. This study shows how powerful genome-scale analyses can be in cancer research.

In addition to the TCGA, we've started the Washington University Cancer Genome Initiative. The project started in April of 2009 and our goal is to sequence tumor tissue and normal tissue from 150 patients in 12 months. We'd like to increase efficiency and reduce costs so that by the end of the project, we can sequence and analyze both sets of tissues for $50,000. I expect that personal genomics will be used to diagnose disease and design treatments very soon.

L: $50,000 still sounds expensive to me! Do you think personal genomics will ever get much cheaper than that?

RW: Sure. Although the next generation sequencing technologies are working very well now, other technologies are also being developed. For example, another generation of sequencers is being designed and tested that can sequence single molecules of DNA with no amplification step.

It won't be too long before genome sequencing plays a major role in diagnosis and treatment of human disease, so we're going to need to start thinking now about how we're going to deal with that information. A law has recently been passed to help safeguard genetic information. It's called GINA, the Genetic Information Nondiscrimination Act. This act was passed by Congress and signed into law by the President in 2008. It prohibits the improper use of genetic information by health insurance companies and employers.

Many in the sequencing community are working towards a goal to sequence a human genome more quickly, more cheaply, and more accurately. I think we can get there soon.

L: It sounds like you're well on your way. Thanks for taking the time to talk to me, Dr. Wilson.

RW: You're welcome, Libby. I hope you enjoyed your visit to the Genome Center.

Additional resources/websites:

Genome Center

http://genome.wustl.edu

Human microbiome project

http://nihroadmap.nih.gov/hmp

Human Gut Microbiome Initiative

http://genome.wustl.edu/projects/human_gut_microbiome_initiative

http://videonews.wustl.edu/?play=Human_microbiome

1000 genomes

http://www.1000genomes.org/page.php

The Cancer Genome Atlas

http://cancergenome.nih.gov/

References:

Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. Oct 23;455(7216):1061-8.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. Nov 6;456(7218):66-72.

Mardis ER. 2006 Anticipating the 1,000 dollar genome. *Genome Biol*. 7(7):112.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.*;9:387-402.

Mardis ER. 2009. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med.* Apr 17;1(4):40.

Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD, Fulton LA, Locke DP, Magrini VJ, Abbott RM, Vickery TL, Reed JS, Robinson JS, Wylie T, Smith SM, Carmichael L, Eldred JM, Harris CC, Walker J, Peck JB, Du F, Dukes AF, Sanderson GE, Brummett AM, Clark E, McMichael JF, Meyer RJ, Schindler JK, Pohl CS, Wallis JW, Shi X, Lin L, Schmidt H, Tang Y, Haipek C, Wiechert ME, Ivy JV, Kalicki J, Elliott G, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson MA, Baty J, Heath S, Shannon WD, Nagarajan R, Link DC, Walter MJ, Graubert TA, DiPersio JF, Wilson RK, Ley TJ. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med.* Sep 10;361(11):1058-66. Epub 2009 Aug 5.

Rothberg JM, Leamon JH. 2008. The development and impact of 454 sequencing. *Nat Biotechnol*. Oct;26(10):1117-24.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature*. Oct 18;449(7164):804-10.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. *Nature*. Jan 22;457(7228):480-4.

Turnbaugh PJ, Gordon JI. 2009. The core gut microbiome, energy balance, and obesity. *J Physiol*. 2009 Jun 2. [Epub ahead of print]

Walter MJ, Payton JE, Ries RE, Shannon WD, Deshmukh H, Zhao Y, Baty J, Heath S, Westervelt P, Watson MA, Tomasson MH, Nagarajan R, O'Gara BP, Bloomfield CD, Mrózek K, Selzer RR, Richmond TA, Kitzman J, Geoghegan J, Eis PS, Maupin R, Fulton RS, McLellan M, Wilson RK, Mardis ER, Link DC, Graubert TA, DiPersio JF, Ley TJ. 2009. Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci*. Aug 4;106(31):12950-5. Epub 2009 Jul 27.