ASIC200 PERSONAL GENOMICS LECTURE (IN CLASS)

(Note, for exam, specific details outlined in the class lecture notes will not be tested on. However, exam questions may involve queries that recognize your understanding of general trends in genetic technology development)

(Building off from the last video)

FYI: Sanger sequencing is good for getting about 1000 nucleotides of sequencing in a single reaction (<- which is kind of tiny if you want to sequence a genome!)

ANYWAY sequencing has allowed us to figure out some pretty remarkable stuff, such as the general structure of what a human gene might look, but more importantly, it has allowed us (from about the 1960s to the present day), to figure out some very fine tuned details about specific genes/proteins. i.e. we know a fair bit about how genes work, how the DNA code is organized, how genes are turned on, how they are turned off, how a single gene can actually have multiple roles.

BASICALLY, the research pipeline worked in this way for a while. Whereby, you have:

1) an interesting organism,
2) with an intriguing phenotype which you try to functionally make sense of,
3) which is linked to gene or two (or more)
4) which in turn is sequenced so that gain info how that gene might work exactly.

And in this way, a single organism (such as a human) gets "figured out" one gene at a time, until you have a sense of how it all works together. i.e. you get a sense of how all these genes might work in the context of the whole organism, in the context of the whole DNA code. NOTE that this is information culled from a variety of sources, not a single specimen. i.e. the data (from many samples) is representative of a species, but the data is not necessarily derived from a single sample of species.

ALSO NOTE that the data has an added layer of complexity, because genes don't tend to be an ALL or NONE thing. You have genes that may exist in forms (ALLELES) that work better, work slower, only work when such and such is just right. Everyone is different right?

BOTTOM line, this mass of data, whilst useful, has many caveats. It represents a holistic "average" of what an organism is all about, and in many ways the only things that get studied are the things which come with very noticeable phenotypes (i.e. lots of stuff just gets missed out – think about how your Facebook profile doesn't quite present a full picture of you!

BECAUSE OF THIS, THE NOTION OF SEQUENCING AN ENTIRE GENOME SORT OF MAKES SENSE, AND THEREFORE PROVIDED THE NEXT STEP. To discuss the implications of this, the drama of the human genome project is probably the best avenue of exploration.

THE (HUMAN) GENOME

First, a reminder of some definitions:

GENOME: In modern molecular biology and genetics, the **genome** is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and the non-coding sequences of the DNA/RNA

GENOMICS: is a discipline in genetics concerning the study of the genomes of organisms.

HUMAN GENOME PROJECT (completed June 26th, 2000): I'll avoid the dramatic bits (which I have a feeling may be discussed by Allen). Briefly…

PUBLIC: In the U.S. most of the funds come from the National Institutes of Health – refered to as the "public project" about $3 billion dollars set over 15years.
Dr. Francis Collins, director of the National Human Genome Research

(started in 1990) "The Human Genome Initiative is a worldwide research effort with the goal of analyzing the structure of human DNA and determining the location of the estimated 100,000 human genes..." initial draft proposal for NIH funding of human genome project.
~30 different human cell libraries.

FOR-PROFIT: most famous enterprise is Celera (Latin for "speed") Genomics Group.

SPEED MATTERS: J. Craig Venter, and NIH scientist frustrated at the slow pace of sequencing the genome leaves NIH in 1992, to form The Institute for Genomic Research (TIGR) with wife Claire Frasier. Venter and others developed a technique termed SHOTGUN SEQUENCING which relies heavily on automated DNA sequencing machines, and in 1995 are the first group in the world to sequence a complete genome (Haemophilus influenzae, a bacteria that causes the flu)

1998 Venter teamed up with PE Biosystems / Applied Biosystems (ABI) to form Celera. Goal sequence the human genome by 2001 (2 years before completion by the HGP, and for a mere $300 million (about a tenth of the public project).

Sample derived from DNA samples taken from 5 individuals. Celera says it used one man's DNA as the foundation for its work. (This turned out to be Dr. Venter himself)

In the end, both project streams agreed to share data to help complete the human genome sequence by mid 2000. This was approximately sequence that worked out to about 10 fold coverage (i.e. actually tried to sequence the sucker 10 times!) for 90% of the total sequence.

SOME KEY FINDINGS

• The HGP has revealed that there are probably about 25,000 to 40,000 (since updated to a count of ~20,500 human genes)
This is really quite amazing! That something as complicated as a living organism such as a human can be derived from the actions and mechanisms of this few working parts! Point in comparison is to realize that total number of different lego pieces (as of May 2010) was about 21,000.

• Human genome is remarkably similar to other genomes in terms of total gene humbers and gene functions, although most genes are more complex. (Comparitive Genomics)
This, folks, was already sort of known but the HGP really segmented this view. That there is value in working with other simpler organisms to pose biological questions that may be applicable to human biology. Use e.coli as an example, in that e.coli is just way less maintenance to look after and study, than say a human.

• Between 1.1% to 1.4% of the genome's sequence codes for proteins. Nonfunctional regions appear to account for ~97%. 12% of human genomic DNA is due to copy number variations – CNVs
The main thing here is to recognize that an awful lot (~97% of the DNA sequence) appears to be completely useless

• ~2 million single nucleotide polymorphisms - SNPs (~0.1 to 0.3% of total genome)
SNPs are pretty important. They deserve their own section!

LOOKING AT SNPs

A single nucleotide polymorphism is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between (human) members.

SNPs are extremely useful because they are a significant element that differentiates one human genome from another. In other words, if we want to sort out why genetic differences result in differences between people, there is (conceptually) no longer a need to sequence the entire genome. JUST LOOK AT SNPs, since they represent a key part of what's different between genomes.

Plus, there's a really cool way to look at SNPs, actually millions of them (at once).

First, a quick rehash of the structure of DNA.
"The double helix structure of DNA is not unlike square dancing in an overly homophobic community"

The double stranded nature of the DNA molecule is actually very useful, because it infers the notion that if you know the sequence of one strand, you can predict the sequence of the other strand (due to the complementation of different nucleotides – A pair with Ts, Gs with Cs)

Powerful because you can look for interactions between complementary sequences.

FOR EXAMPLE: (in class we'll reenact by using the square dancing situation as our metaphor for SNPs.) Let's say we are interested in finding out whether a person's genome has 3 specific SNPs (i.e. specific single nucleotide polymorphisms).

If we have a single strand of code that denotes the SNP#1 and its surrounding sequences, we can attach that to some sort of solid support. Then we can do the same for the other SNPs (SNP#2 to SNP#3).

This might look a little like this:

|•ATC (SNP = I hate Star Wars)

|•TTG (SNP = I hate ASIC200)

|•CAG (SNP = I hate Nickelback)

Now if we take a sample DNA -> (1) break it up into smaller pieces, (2) make it single stranded (you can do this by heating DNA); and also (3) label it somehow, such as attach a fluorescent or glow in dark tag. This might look a little like this.

GGTCGAATGCGATTTC (initial sample)

GGT CGA ATG AAC TTTC (broken up)

GGT^ CGA^ ATG^ **AAC^** TTTC^ (labeled with ^)

Noticed that the blue sequence is complementary to the blue sequence attached to the solid support. This means that these two will bind specifically, culminating in the ^ label positioning itself in the exact same spatial place as where the TTG strand was placed.

i.e. because we see a signal at the middle sequence (due to the binding), we can infer that "Hey, this sample we've got must have sequences that indicate the presence of SNP#2.

Of course, in this example we're only looking at 3 SNPs (and in reality these probes are not just 3 nucleotides in length, they tend to be around the 20 nucleotide length). But what about 100 SNPs, or maybe a thousand? Or let's say even a million IN ONE EXPERIMENT. If you can do a look for binding with a million different sequences, then you've got yourself a pretty powerful system of looking at SNPs. And remember, SNPs represent a significant portion of what makes one person's genome different from another!

Anyway, all to say that such experiments – looking at millions of SNPs in one go is indeed possible. We call these things DNA CHIPS or MICROARRAYS.

As well, this is a powerful form of GENOTYPING.

(definition) **GENOTYPING**: is the process of determining information about the genes (genotype) of an individual by examining the individual's DNA sequence by using biological experiments (such as looking for SNP pairing)

The ability to genotype in this way is actually very powerful (use horse vs unicorn example again). For instance, if a SNP is well defined – i.e. if you have this SNP, then that means you have this trait – then you can use this for predictions.

However, it also allows you to more quickly and efficiently infer linkage between a trait and DNA sequence, sometimes simply by looking at differences or similarities in DNA sequences and attempting to correlate trait differences and or similarities. (USE UNICORN vs HORSE example). i.e. if all unicorns have this SNP, and all horses are missing the SNP, then maybe the SNP has something to do with unicornism?

NOTE that this is looking for a correlation trend, and correlation DOES NOT equal CAUSATION but if the correlation is very striking, and spans over a massive sample number, then it's probably going to be interesting enough for you to want to check it out further.

ALSO: Bring up the HAPMAP project (public project to characterize all possible human SNPs – up to about 10 million so far), as well as services like "23andme" (which recently got halted due to an FDA ruling – Allen will get into this – this is the $99 genotyping service (which essentially looks at a variety of medically relevant SNPs).

BETTER WAYS TO JUST SEQUENCE THE HECK OUT OF A SINGLE SAMPLE

Although looking at SNPs is a powerful way of quickly characterizing a large number of elements in a person's genome, it stands to reason that if you could just sequence the **whole** genome of many individuals, perhaps even all individuals, then you would have an even stronger data set from which to correlate (and therefore identify) DNA sequences that result in certain traits/phenotypes.

I.e. instead of each sample being represented by 2 million SNPs, each sample is instead represented by 3 billion nucleotides! Obviously, you need some serious computer power to be able to look at this effectively – but guess what? Computers are already there.

Anyway, it's already happening. The first full genomes were sequenced and published in 2007. Craig Venter and James Watson. There's also the completed **1000 genome project**, which had 1000 human genomes (basically volunteers) sequenced by 2012, where all public data in the hopes of being able to correlate phenotypes/traits of these 1000 individuals to their genetic code.

And sequencing technology is getting better all the time. In fact, the speed in which sequencing is increasing is often compared to *Moore's Law* (law of transistors – whereby The quantity of

transistors that can be placed inexpensively on an integrated circuit has doubled approximately every two years).

From the two graphs, you can see how sequencing technology is improving from both an output point of view (how many letters we can get and fast it takes to get it), as well as a cost per Mbp point of view. (Mbp = 1000,000bp, letters). It's actually improving FASTER than Moore's Law!

HOW IS THIS POSSIBLE? Various new technologies that allow for this. Go back to the Sanger example. With that experiment, you can get about 1000 letters in a single experiment, which will take about a full day. However, that is one tube, one experiment – you can imagine, it's pretty straightforward to (say) work with 10 tubes. This means, that in a single day, you can get 10x1000letters of data = 10,000bp. Conceptually, this means I can increase my sequencing output if I just have the opportunity to do more reactions in a single go.

However, this also raises costs – i.e. instead of chemicals for one tube, I need chemicals for 10 tubes (or however many reactions I do, since they are all in separate tubes)
MEANING that the challenge one has (in increasing the amount of code determined, as well as keeping costs down) is: Can I mimic the data of millions of reactions, but all in one place (one tube). This is what we'll go over with some more technical detail next week (i.e. it's very cool)

Anyway, THESE are what the new technologies are all about! (I'll highlight Illumina's Solexa platform next week).

THE NET RESULT, is that research is now at a place where getting LOTS of sequencing data for relatively cheap costs is doable. Here are some graphs to show you where we're at now.

GENBANK graph of Gb of sequencing data (per year). Now, see what machines such as Illumina's Solexis technology can do!

Two Illumina MySEQ illumine bench top machines working for one year, can generate ~400Gigabases of data. This is roughly double the totality of what was sequenced from 1990 to 2007.

Another striking statistic: Vancouver's Genome Science Center. (from Marco Marra)
4 trillion bp (1999 – 2010)
173 trillion bp (2011 to April 2012)

In any event, currently biological research is being propelled by these technologies, because ultimately, they allow you to get massive amounts of raw data (DNA, or RNA code). Much like google algorithms, the trick is correlate this raw data with phenotypic observation, using computer tools, and use statistics to access potential validity.

BUT, when your data sample consists of 3 billion variables, it's actually not that difficult to find correlations that are irrevelant, i.e. false correlations or the simple fact that if you're looking at 3 billion different things then of course, you're gonna find correlations, but of course, most of them would be statistically coincidental.

(See Unicorn versus Horse example)

How do you fix this problem with false correlations. Well you need to sequence more genomes, so that you have more samples to look for correlations. However, the issue with this, is that as fast and cheap as sequencing is, it was still too expensive to do at a scale that statistically approaches numbers of samples needed.

But this speed and cost is changing: and fast. And with that, I'll leave you with the following stats so that you can compare how things have changed in even the last few years, we've taught this course…

AT LAUNCH OF HUMAN GENOME PROJECT (1990)
Several machines to sequence the human genome. Est. time and cost: 15 years and $3 billion

4 years ago (2012):
One machine can sequence an entire genome in about 8 days at a cost of about $10,000

3 year ago (2013):
One machine can sequence an entire genome in about 3 days at a cost of about $5,000

2 years ago (2012):
A suite of high powered machines can sequence an entire genome in about a day for about $1000.

October 2015:
One machine can sequence an entire genome in about 1 day at a cost of about $1200.

CRISPR/Cas

It's here folks – easy, cheap, genome editing.

Essentially, what you have here is a bacteria immune system that was discovered in the late 80s where incoming foreign (and hence bad) DNA from viruses got into bacteria to wreck havoc. However, it was noted that some of this foreign viral DNA was chopped up, and then integrated into special areas of the bacteria's own genome (these are CRISPR areas). In a way, it was a mechanism for the bacteria to "remember" these viral sequences, since they are being kept in these CRISPR areas.

From there, it was discovered that another bacterial protein (Cas9) was capable of holding on to these "saved" viral sequences, and use them as homing devices to cause DNA cutting. i.e. Cas wants to cut DNA, but only cuts DNA sequences that matches the viral sequences – basically, a proto memory Immune system!

Anyway, what folks found is that this CRISPR/Cas system works really really really well in other cell types (most notably mammalian systems), which has made this genome editing tool all the rage in recent years. So much so, that there's a brutal patent war going on (lots of money at stake), as well as various declared moratoriums on the use of the system in human cells (in essence, because it's a little too easy to use).