ASIC200 Notes: Personal Genomics: Video (@ng_dave)

- - -

VIDEO ONE

0.1 CONTEXT

At the end of the day, and whether we realize it or not, we are concerned about the living. Whether this is in the context of improving our health or the health of those around us, understanding the players in our ecosystems, or obtaining resources for food and/or materials, many of the advantages that our society have, are a result of our ability to understand and harness the living. And broadly speaking, this means that we want to know as much as we possibly can: about our own human bodies, and also biodiversity at large. In fact, we see value in not just understanding it, but in perhaps being able to predict it, or even control it.

Genomics, and its sister terms (proteomics, metabolomics, etc) are one avenue of such exploration. What makes them powerful, however, is that they operate under a flushness of information. – the biggest pile of data you can imagine, all of which exists in molecular parts, if not in digital form.

Personal genomics is a branch of science where individual genomes are analyzed and characterized using computer tools (this is the main conceptual topic in which we'll apply these notes to).

It's also very powerful, quite amazing, if not a little scary sometimes. But that's why it's good to have a clear handle on the science involved.

- - -

Let's start with a small glossary...

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

DEOXYRIBONUCLEIC ACID or **DNA** is a molecule that contains the genetic instructions used in the development and functioning of (almost) all known living organisms.

NUCLEOTIDES are molecules that, when joined together, make up the structural units of DNA.

A **GENE** is a unit of heredity in a living organism. It normally resides on some stretches of DNA and RNA that codes for a type **PROTEIN** that has a FUNCTION in the organism.

The FUNCTION of a GENE PRODUCT can (by itself or in tandem with other GENE PRODUCTS) result in an observable **PHENOTYPE** or **TRAIT**.

An **ALLELE** is one of two or more forms of a gene or a genetic locus (generally a group of genes). Sometimes, different alleles can result in different observable phenotypic traits, such as different pigmentation. However, many variations at the genetic level result in little or no observable variation. Think of an **ALLELE** as a specific version of a **GENE**.

0.2 THE DNA (BASICS)

From most perspectives, a lot of people would feel that the basic mode of data that scientists have been using is DNA. This is sort of true, and for now, close enough for us to dig a little deeper.

- - -

So what exactly are the key features of an organisms DNA? Well, central to this is the idea that the DNA contained within an organism (the genome) is analogous to a blueprint for the construction and operation of that organism. In other words, the DNA is very much like an instruction manual – a very detailed and voluminous instruction manual.

No offense to all the wonderfully talented individuals in the world, but Mother Nature has really outdone herself here with a rather superb job of getting this genome business to work. It is nothing short of amazing.

This instruction manual is basically a code that is written in the language of a molecule called *deoxyribonucleic acid*, (this here is our *DNA)*. DNA is this rather pretty looking molecule that is composed of four different building blocks. Together, these building blocks are known as *nucleotides*, and individually they each have a chemical name which is often abbreviated with a single letter - these letters being A for *adenine*, T for *thymine*, C for *cytosine*, and G for *guanine*. In effect, your DNA code is much like a language, a script of sorts, with the principle difference being that it is composed of only four letters instead of the full twenty six.

As alluded to earlier, a classic example of what your DNA code is capable of doing is the textbook case of natural eye colour. Your eyes are a certain colour because of the instructions within your DNA. The same is also true for your natural hair colour, and in other school examples such as whether you are able to roll your tongue or not. However, it's important to realize that virtually *every* physical attribute you have is determined by your genetic makeup. In other words, this also includes subtle nuances like the fact that some of your acquaintances are more prone to farting when they ingest dairy products or that a few of your friends may get drunk more quickly than others. Taken together, this means that your DNA is responsible for an awful lot of information, which at first glimpse is difficult to fully appreciate.

To put this all in perspective, it's important to try and visualize the enormity of the task at hand. One good way of doing this is to concentrate and focus on one of your thumbs. Ask yourself a few simple questions. How does your thumb know that it's a thumb? How does its cells distinguish themselves from the cells of other fingers? How does it know to come out of a certain place on your hand – next to your forefinger, not next to the pinkie finger? For that matter, how does it even know that it should be protruding from your hand and not from your foot? In truth, these somewhat bizarre thoughts centre round a field of research known as developmental biology. These sorts of scientific questions are constantly asked in this dynamic field, although not necessarily always for the thumbs - rather for the architecture of the entire body, or even possibly some other creature's body. In essence, these biologists continually think about the following question. How do we go from a single cell entity, created from a marriage of a sperm and an egg, to a being of very set features, full of many different types of cells and many different types of tissues? If you look around you, distinct though we are from each other, we are all basically the same. What I mean here is that generally speaking (and I hope I don't offend anyone), we all have heads, we all have torsos, and so on and so on. Furthermore, all of these bits and pieces are usually in the right sorts of places.

You must remember that it is your genome that is providing and directing all of this information. Imagine doing this yourself with pen and paper. Think of all the countless notes and scribbles you would need, so that something as basic as your body shape is done properly. For example, you may need to devote a few pages to your eyebrows. You would need to ensure that your eyebrows are in the right place. Not anywhere unsettling like your nipples, but somewhere on

your face. Over your eyes and not under your eyes. You would need to describe their thickness, their length, and their colour. Hopefully, you get the point - the details are endless. Simply put, the amount of physical information in your DNA code is mind boggling.

And it doesn't even stop there. Although a bit more controversial, it is becoming clear that an individual's general behaviour and personality is, in part, predetermined by your DNA. Obviously in this case, a person's environment and experience plays a vastly more dominant role, but there is nevertheless plenty of evidence to suggest the importance of genetic factors in these types of traits. (Mention dog genomics as an example).

SOME NUMBERS
A fairly good estimate of the size of your human genome is a total of 3.0 billion letters of code.

Other genomes: ~390,000,000 bp (rice), ~2,400,000,000 bp (dog), 4,639,221 bp (e.coli), 157,000,000bp (*Arabidopsis*), 18,000,000,000 + (*Pinus* tree)

It's worth noting that these are actually really huge numbers, the scale of which I find is often lost to the casual listener. You get habituated, I think, by references to the country being in debt so many billion dollars, or by certain athletes signing billion dollar contracts. This is, matter of fact, a *very* big number and many other analogies abound that are much more eloquent than my example. If we were, for instance, to take 3.0 billion nucleotides and translate them into text, letter for letter, the genome in its entirety would be equivalent to about 8000-9000 copies of the first Harry Potter book. Another one is to take 3.0 billion grains of sugar and pile them up in one spot. Apart from wasting a lot of sugar, you would discover that you would've formed a mound about the size of three cars. And we could even say that piling 3.0 billion cars on top of each other would likely resemble a mountain of Everest proportions.

Regardless of the analogy you use, the shear size of your genome does make sense. It is, after all, responsible for so many things, and you would assume that you would need that much information to get all the details and all the nuances sorted out.

ALSO note that actual “letters” of code per person can be defined by other ways.

i.e. DNA is double stranded. Also that organisms can have multiple copies within their own cells. An obvious example of this is that humans are DIPLOID (meaning that our code technically is a dual set. One from your Mother and one from your Father).

VIDEO TWO

0.3 CODE = PHENOTYPE

How exactly does DNA code translate itself into an observable characteristic, a phenotype?

Here, we need to go over the CENTRAL DOGMA. Essentially a mechanism of going from “instruction booklet” to an actual “product.”

For now, ignore the bit about RNA

...Which is very much about proteins: in fact Martha Stewart would say, "proteins are a good thing." In life, they are the true movers and shakers of any organism, and are the molecules that

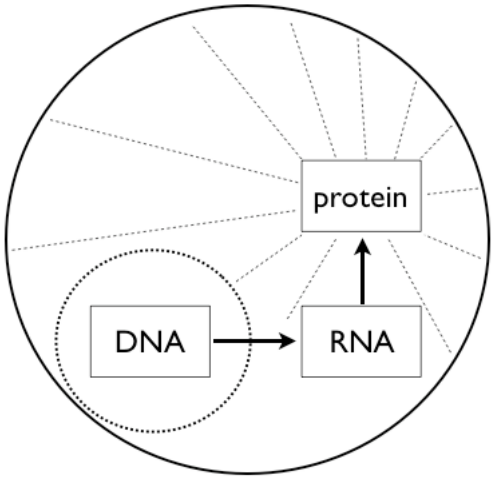
actually go through the daily business of living. In short, these are the building blocks that give your various tissues their shape and their function. I bring this up now, because all of this talk about genomes and DNA is only illuminating if you recognize the fact that your genetic code is simply the instructional package for making all of the different types of proteins needed for life.

And things you need for life include: proteins that regulate chemical reactions (for instance, in the conversion of the food we eat into energy); proteins that transport key molecules from one place to another (like the pump implicated in cystic fibrosis); proteins that become the basis of cell structure (like how the architecture of certain tissues is achieved); proteins that facilitate cellular communication (how all the different bits and pieces of your body can work together). In truth, the diversity in protein makeup is responsible for all the diversity in life itself. In other words, bring on the bacon.

In itself, how proteins come about from your DNA code is quite clever. Proteins are built by piecing together molecules that are collectively called *amino acids*. It's a bit analogous to DNA in that if you recall, your DNA code consists of specific combinations of nucleotides. However, whereas our DNA is composed of an alphabet of only four different letters (A, T, C, and G), proteins are built with a much larger alphabet of 20 different letters or 20 different *amino acids.* Again, for any given protein, the determination of which of the 20 amino acids to use and in what order they are to be pieced together is dependant on the nucleotide code itself. This might sound a little confusing but in essence the production of proteins is dependent on dealing with two types of code. More specifically, each combination of *three* nucleotides (often referred to as a *codon*) will signify a particular amino acid. For example, the nucleotide T, followed by a G and another G (or TGG) codes for the amino acid Tryptophan (abbreviated ‘W’), the sequence ATG codes for the amino acid Methionine (abbreviated ‘M’), and so on. The sequence TGGATG would therefore code for two adjacent amino acids, Tryptophan and Methionine. Altogether, there are three letter codons for all 20 different amino acids. In this manner, a long sequence of nucleotides can potentially and theoretically be translated into a long chain of amino acids - i.e. a protein molecule.

To illustrate this two code system, the best example that comes to mind, is the use of Morse code to send messages overseas. In this situation, you essentially have a binary code (dot or dash, two options), which when rearranged into units of three, can translate into one of the 26 letters. For example, *dot dot dot* is the same as the letter 'S', and *dash dash dash* is the same as the letter 'O'. This is a two code system. Your first part being the Morse code element, and the second part being the formation of words from letters. In our biological example, the first code involves the use of nucleotides to provide information for which amino acids to use, whereas the second code dictates the length and combination of amino acids to form a functionally relevant protein.

(Now let’s take back the bit about ignoring the RNA)
In truth, the relationship between proteins and DNA is a little bit more complicated. First, it turns out that the overwhelming majority of the human genome doesn't do anything, and is basically considered to be garbage, junk, filler or if you want to be particularly nasty, crap. This accounts for an astounding 97% of your genome having absolutely no function or significance. This introduces an interesting logistical problem in that humans are using what is essentially a polluted genetic code. In other words, there has to be a system that allows the deciphering of the good stuff from the bad. You don't want to waste your time decoding your junk regions in that it could translate into some random, useless or potentially harmful protein.



Even your freakin' earlobe is doing this right now!

Secondly, the location of your DNA and the location of protein synthesis are different. This of course, makes no sense because how can you translate your DNA into proteins if the two molecules reside in geographically distinct places? Here, we find that your DNA is found within a small physically enclosed area of the cell called a *nucleus*, and proteins are awkwardly made *outside the nucleus*. Although this nucleus could be viewed as simply a mechanism to "house" and protect your genomic DNA, it does create a rather unfortunate conundrum in that the all important DNA code is not accessible to the machinery necessary for its translation into proteins.

In a rather crafty way, biology has managed to solve these problems through the use of a middle-man known as the *messenger RNA* molecule or *mRNA* for short. For the sake of clarity, mRNA is fundamentally similar in structure to DNA having nucleotides. There is a slight difference but it's visually quite minor - it could actually be the basis of a challenging 'spot the difference' comic. However, thinking conceptually, mRNA is comparable to a no-nonsense piece of genetic code that is constructed from only the useful parts of your DNA. This is similar to having study notes for a particular subject where only the crucial parts are highlighted and regurgitated. Consequently, problem one is solved. Here we have a strategy that can weed out the good from the bad and hence no crap.

Additionally, mRNA is special in that it is a string of nucleotides with the ability to move and ultimately leave the nucleus. You have to remember that your genomic DNA living inside the nucleus of a cell is akin to an elephant stuck in the upstairs toilet. It is simply too big to pass through doors that might otherwise be situated along the walls of the nucleus. mRNA molecules do not need to be so big. They are much more manageable in size because for each molecule, they contain only the sequence of one protein (not all of them), and more importantly they contain only the *necessary* sequence of that one protein (no junk). This means that problem two is also solved, as mRNA acts as a mobile representative of the genetic code that can now get out and come into contact with components required for protein production.

- - -

Confused? Don't worry, it's alright if it seems a little perplexing right now. I know many people who have had nightmares over this stuff. If you do find yourself waking up in the middle of the night screaming nonsense about RNA and elephants, try thinking of the following analogy.

Because you are such a wonderful person, you wish to prepare a nice chocolate cake for your friend, and to do this, you visit the library to look for a good cake recipe. For some unexplained reason, you are also a huge Martha Stewart fan, which is why you decide to look for a cake recipe in one of her many 'Martha Stewart Living' magazines. After searching for several hours, lo and behold, you find a promising recipe in her 'Weddings Issue', but notice that the magazine itself has a sticker on it that says 'for reference only.' This is a bit of a bother because it means that you won't be able to take the magazine out of the library, and hence, into your kitchen where you had plan to spend most of your time being a wonderful person. Furthermore, despite your best efforts, you can't seem to find any semblance of a photocopy machine anywhere, since this is the sort of library this is, and since the analogy wouldn't work otherwise. Begrudgingly, this small nuisance forces you to look for a pen and a piece of paper so that you can manually scribble down the recipe to take home. As you do this, you quietly think to yourself that your friend had better appreciate all of this effort.

No offence to Ms. Stewart, but I find her magazines are always full of extraneous and in my opinion useless information. Do you really need to know the history of the chocolate cake? Do you really need to know about the appropriate cutlery used for serving cake? Do you really need to see and evaluate 15 different colour schemes for acceptable presentation? I don't think so. All you really need to concern yourself with is the ability to make the cake taste good. This is why, when you go to the bother of copying down the recipe, you don't include all of the nonsense - you just copy down what you need. In short, this turns out to be just a few lines of ingredients and directions scrawled neatly on your piece of paper. The crucial point is that you can now freely walk out of the library with the recipe in hand.

Next, of course, is a trip to the local grocery store where you would get all the necessary cake ingredients and maybe indulge yourself with the smutty magazine about child actors gone bad. After which, you would head home and bake a wonderful chocolate cake which is met with such praise, that you are glad you didn't waste your time using table setting number six for the occasion.

A strange story indeed but here is how the analogy works. First, you need to envision the entire library with all of its resources as the genome, and also envision the building itself as the nucleus. The complete recipe found in the magazine actually represents the genomic sequence for one particular protein. As mentioned before, Martha publications tend to have a lot of useless information, some of which is not even directly related to the production of the cake (advertisements and historic footnotes). This is identical in premise to the crap in your genomic material, and the concise notes you scribbled down symbolize the messenger RNA molecule. This, as mentioned before, is twofold important because, (1) it represents the minimal amount of information needed and, (2) it represents the ability to leave the nucleus (in this case, the library) and the ability to go to places where protein production can take place (in this case, the rest of the world, but more importantly the grocery store and your kitchen). Finally, the cake itself represents the protein. Remember I said that in living systems, it's really the proteins that are the real movers and shakers? They are certainly the most interesting parts of the big biological picture, and wouldn't you say that the cake itself is the most interesting part of this process?

ANYWAY, this is some of the basics of genetics (replication, central dogman, DNA, RNA, protein, genes, genome). This stuff is really powerful, and clocking along at a phenomenal speed.

Anyway, consider the following:

ONE: That traits and phenotypes are what we pragmatically care about.

TWO: That these are reflected by a variety of different types of molecules.

THREE: this means that we technically have different places to “look.” i.e. DNA, or RNA, or Protein.

Taken together, this is why we have the field of molecular biology. This is essentially a catch phrase that encompasses the science and methodologies that allows us to look at these molecular components, especially ones that divulge some information about how the code (at various stages) becomes the phenotype.

VIDEO THREE

MOSTLY REPLICATION
Here, you must read and study the following piece on replication. Please go to http://popperfont.net/2011/09/05/breakfast-of-champions-does-replication/ If you don't the video about sequencing, as well as what we'll be doing in the lab will not make any sense!

VIDEO FOUR

0.4 MOLECULAR BIOLOGY

Looking at the molecules involved in biological processes! i.e. How to study DNA, RNA, and proteins!

GEL ELECTROPHORESIS
Experiment that creates a gel composed of a filamentous material (Agarose in the case of DNA gels). Essentially, these molecules crisscross and form a net like network. The DNA is then forced through this mesh by way of an electric field (i.e. one side of the gel is charged, and that attracts the DNA to run through in a certain direction). Basically, the bigger you are, the more you are impeded by the mesh and the slower you run. Therefore size if based on speed and distance migrated through the mesh where small = fast, and large = slow.

(Use Hamster analogy).

Will also cover the act of "sequencing DNA." The conventional technique for this (which is still widely used) is known as SANGER SEQUENCING.

SANGER SEQUENCING (Not required for exam)

Often known as the Chain Termination procedure – technical details are essentially a replication process in the presence of dideoxynucleotides (nucleotides missing their 3'OH ends). Since (according to polymerase rules), this is the end required for chain elongation, missing this 3' end automatically results in replication stopping.

Since the stoppage is happening at a place dependant on the specific dideoynucleotide used, the size of the fragments you generate can allude to the sequence of the original template (watch the video for full details).

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!)