**A set of notes by Charlie** **Spring 2017**

# Notes on an 'Introduction to Genomic Technologies'

*First Module of the Coursera Genomic Data Science Specialization from John Hopkins University*

## Week One: Overview

Why study genomics? Everybody has a genome which governs their biology. Despite differences in peoples appearances, we are 99.9% identical. But what drives the differences? We start as a single cell, dividing into an embryo, and a whole person: all encoded within our genome. This code determines all the different cell types. Another big area of research in genomics is cancer: cells in your body which replicate without control. Mutations are changes in genome because of damage or errors in replication.

**The central dogma of molecular biology: information flows in a single direction from your genome (DNA) to RNA to proteins. DNA contains the code for making proteins - mRNA is a copy located on the DNA molecule which leaves the nucleus of the cell, and the ribosome reads the coding sequence to put amino acids together.**

Proteins are comprised of 20 letters (amino acids) – long molecules (3-400 amino acids long). Made from triplets on RNA, where each triplet encodes an amino acid. There are four possible RNA nucleotides, so there are $3^4$ (64) triplets which get translated either into an amino acid or one of 3 stop codons which end a sequence. Over time, we've learnt that information can flow both ways: some proteins bind and modify the DNA - self regulating - and the dogma has been largely refuted. Other modifiers can affect DNA itself. Sequencing technology is at the heart of the genomics revolution: creating enormous datasets which now take longer to analyze than create. The cost of sequencing a genome has dropped from $100m to about $10k since about 2002. The biggest international repository is NCBI.

Genomics is the branch of molecular biology concerned with the structure, function, evolution and mapping of genomes – where a genome is all of the molecular information inside your cells which defines how your body works. The structure is a sequence of four nucleotides (A,C,G,T), and in the human genome there are 3bn of these letters strung together, divided into 23 chromozone pairs – 22 identical and then {x,y} for men, and {x,x} for women. Within each chromosone there's a centromere, and at the end, a telomore. A chromozone is a thread-like structure of nucleic acids and protein found in the nucleus of most living cells, carrying genetic information in the form of genes. The function of a genome is all the things it does: how to develope from an embryo, respiration, metabolism, etc. Chimpanzies diverged from us about 6m years ago – but the sequence similiarities are close, even for unrelated organisms like bacteria. For example: reproduction. A gene is a heritable unit - a small section of the genome which encoded a protein which has some specific function.

**Note: Important difference between genetics and genomics – genetics typically studied one or a few genes, whereas genomics studies all genes at once. Data challenges drive genomics with global, high-throughput experiments.**

Genomic data science is at the intersection of biology, statistics and computer science, also known as computational genomics, computational biology, bioinformatics or statistical genomics. Steps to undertaking genomic data science:

1. Start with 'subjects' - humans, mice, etc. Collect samples from the subjects (e.g. skin cells) – experimental design is important, even though data collection is now cheaper.

2. Prepare the samples in the lab and send them for sequencing, which generates enormous amounts of data.

3. Take sequences ('reads') which are short fragments of the genome - align them to the reference human genome - which represents 'average northern european male' - that tells us how they differ from the reference genome.

4. Preprocess and normalize data in order to correct various types of biases - this is because the sequencing machine can make mistakes (random and non-random), there exist data collection biases, etc.

5. Apply statistics and machine learning techniques to go from preprocessed/normalized data to scientific conclusions.

6. The whole process typically involves a lot of software development.

Moving from individual genomics (i.e. how a cell mutates) to population genomics can show how genomic differences cause recogniseable changes between people. Celluar organisms can be split into 3 domains: 1.) bacteria, 2.) archaea, and 3.) eukaryota - which have cell nuclei, evolved as a way to sequester our DNA from the rest of the cell. Yeast is a single cell eukaryote - which still has a nuclei. Prokaryote (bacteria and archaea) has loosely organized DNA floating around. In Eukarayote cells (surrounded by a wall), there is a nucleus (which has its own wall), inside of which is chromozones (long molecoles of DNA). Mitocondria has its own (small amount of) DNA - the 'powerhouse' of the cell - these genes are responsible for energy metabolism. Over the course of their life, cells have a well defined cycle (e.g. mitosis - DNA replicates within a cell, then the cell splits into two 'diploid' cells). Diploid means that we have two copies of every chromozone - male and female. Not all Eukaryotes are diploid (but humans are). During the course of development - cells need to develop into different type of cells – all begining from stem cells. They go down developmental paths which differentiate them. During sexual reproduction, something different happens: different recombinations - 'crossing over' - part of chromozone 1 from female crosses over into chromosone 2 with male - typically about one crossover per chromozone. This is the main reason why family members aren't necesssarily alike. DNA is the molecule which comprises all of our genetic material, composed of nucleotides of nucleic acids - A,G,C,T ([Adenine, Guannie] - two ring 'Purines', [Cytosine, Thymine (and Uracil)] - Pyrimidines which are a little smaller). A's **always** bind to T's, and G's to C's. This property is very important for how DNA copies itself from one generation to another: even with one strand, you know what the other strand is. Every one of your cells has all of your DNA in it, stored in the famous double helix structure - coiled up and packed into the nucleus. The DNA sequence looks like this - note the direction (3' and 5')!

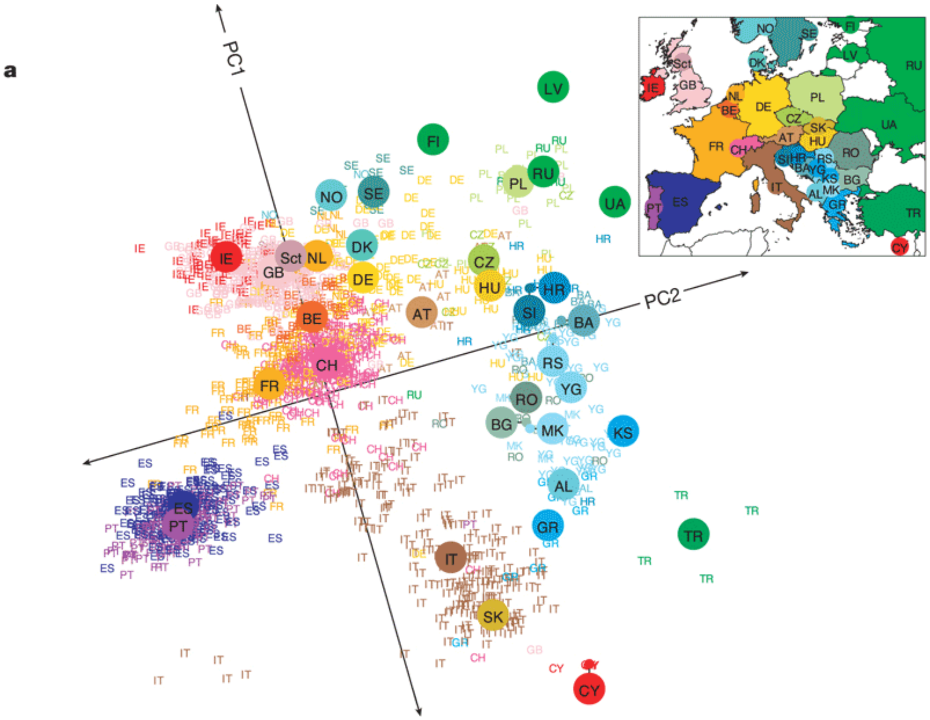
**3' - ACACCGGTT - 5'**
**3' - TGTGGCCAA - 5'**

where the second is the reverse compliament (or 'negative' strand). RNA is *almost* like DNA, the most notable difference is that Thymine is replaced with Uracil - RNA molecules are single stranded in general - and from this RNA template we create proteins. DNAs are replicated between cells, and the RNA uses the DNA template to make proteins (although an otherwise identical template). DNA is basically a program - RNA makes proteins - translate from RNA into proteins. Every combination of 3 letters of RNA encodes an amino acid. Translation machinery reads across RNA molecule 3 at a time - when it hits a codon, thats the end of the protein.

**The Human Genome Project**: first proposed by the Department of Energy in the 1980s, was originally opposed on the basis of cost by geneticists who thought only individual cells should be targetted. In the early 1990s, the plan was to create 'maps' - small peices of DNA and place them on a genome. In 1995, TIGR sequenced the first complete bacterial genome - haemophilus influenza - 1.8 million bases, 1742 genes. In 1998, Applied Biosystems developd a new sequencing machine - faster. Celera Genomics - a for-profit company enters the race, and the public sector increases its efforts in 1998. Each genome is different - the Human Genome Project sequenced one 'mosasic' of about 12 people of northern Eropean descent. Two papers published simultaneously on a draft of the genome - estimates between 30,000-40,000 genes. Now we believe the number is between 20,000 - 22,000. This number relates to the number of protein coding genes - a piece of DNA which gets transcribed into RNA and translated into a protein. However, there are also some genes where DNA gets transcribed into RNA, which itself has a function (without getting translated into a protein).

DNA is inside every cell, 23 chromozone pairs - if you stretch it out - it's about 2m long *per cell*. DNA is coiled around histones - highly alkaline proteins found in eukaryotic cell nuclei that package and order the DNA into structural units called nucleosomes. When translated, it needs to unwrap. The structure also repeats itself, classified either as 'tandem repeats' or 'interspersed repeats'. Typical human messenger RNA has several regions: 5' UTR - 'untranslated region at the start and the 3' UTR - the end untranslated regions. After transcription occurs, a long series of A's is added -'poly-a-tail'. In the middle is the 'coding sequence' - CDS.

Decomposition by PCA (Novembre et al., 2008)



Proteins themselves have structure: the amino acids can also form sheets or helixes. There are thousands and thousands of structures of proteins which confer upon them thier structure. Cells behave differently (i.e. skin and blood) - this is because different genes are active in different cells - controlled by proteins which go back and bind to the DNA itself (accelerate or decelarate) - a control mechanism to decide how much of a protein to produce.

Epigenetic means 'beyond genetics' - factors outside the DNA itself which control something about how the cell functions and how genes are expressed. Affected by: development, environmental chemicals, such as drugs or pharmaceuticals or aging or diet. This can lead to: cancer, autoimmune disease, mental disorders and diabetes. For example - methyl groups can tag DNA and activate or repress genes. The binding of epigenetic factors to histone 'tails' alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

Genotype: collection of all sequences of all genes in all cells - determines how cells functions and whether you have certain traits or diseases - *inherited information*. Phenotypes are traits such as height, hair colour, weight or even personality - *something we observe*. We have two copies of every gene. Traits can either be recessive or dominent: if we have two copies of a gene then we have the recessive trait, and if it's a dominent trait: we only need one copy of that gene. Note: from two pairs (mates), there are four possible combinations of traits - this is how genotype affects/confers phenotype. We have now started to map out genetic variations across the world - specific mutations are common to specific parts of the world - either single nucleotide polymorphisms (SNPS) or larger chunks of DNA. This allows us to categorize populations by their genomes:

The figure ('Decomposition by PCA') shows the axes upon which the genes vary the most with all the information reduced to two dimensions (two PCAs with the most variance). It is possible to measure the connection between genotype (.e.g. AA, AG, GG) and phenotype (brown/green/blue eyes), where, for example, HERC2 is the gene for eye colour. Genome Wide Association Studies (GWAS) study large groups of people and the association between single nucleotide polymorphisms or other associations and diseases or traits of interest.

## Reading for Week One (hyperlinks)

- A nice discussion of (the demise of) central dogma.

- A nice discussion of epigenetics and cancer.
- An overview of what DNA is.
- The real cost of sequencing (i.e. downstream analysis, etc.
- The need to combine bioinformatics and medical informatics.
- An overview of what NGS makes possible.

## Week Two - Measurement Technology

*Polymerase Chain Reaction* (PCR) is a way to copy DNA. DNA is always double stranded: e.g. primers such as 5' ACACCGGTTCGTAGAGCAT 3' and 3' TGGRGGCCAAGCATCTCGTA 5'. But how do we do PCR?

1. Melt gently (94 degrees) - so strands fall apart and so the primers wont stick to the DNA and the two strands of the DNA wont stick to each other.
2. When cooling (anneal - 54 degrees), the primers will stick to the DNA before the DNA strands stick together.
3. We need a second mixture (provided by nature) - a copier molecule - DNA polymerase which we can synthesize - enzymes that synthesize DNA molecules from deoxyribonucleotides, the building blocks of DNA.
4. It will find the sites and fill in missing sequences at primers - 'extension' (72 degrees)
5. Repeat this: with each round, we double the amount of DNA we had before. Typically done 30 times to get about 2 billion copies.

*Summary of PCR:* we need DNA to copy, primers, DNA polymerase and lots of As,Cs,Gs and Ts. Melt at 94 degrees, cool to 54, warm to 72, repeat. *Next Generation Sequencing* – NGS. DNA sequencing began with Sanger DNA sequencing, then DNA microarrays, then 2nd generation DNA sequencing (most common in use now), then 3rd generation & single molecule sequencing (2010 - present). Understanding sequencing techniques helps to understand the data to be analyzed.

First, DNA is copied using DNA polymerase, which takes free nucleotides floating around in a cell/synthesized, to copy DNA trying to be synthesized using rules that Gs bind to Cs, As to Ts, etc. NGS takes template DNA, chops it up into small pieces (maybe 1000 bases long), attaches them to a slide and uses PCR to replicate them. We then hit them with a light which shows them in four different colours (they 'fluoresce'): add them to the slide and have them basepair with nucleotides. We then cycle through the bases. However, errors increase later in cycles: clusters can be 'on schedule', 'behind' or 'ahead' - where bases arent added in the correct sequence. The base calling software also estimates how likely there is an error at a speciic point based on how pure the colour signal was.

The basic idea of NGS applications: convert a molecule to DNA and apply 2nd generation sequencing to measure it (i.e. exon sequencing). Exons are part of RNA - concatenated together to get translated into proteins. For this reason, we only look at the protein coding exons using 'beads' - which hybridize and attaches to exomes. Another technology is RNA sequencing - capture all genes being turned on/expressed in a cell/collection of cells. After transcription, the cell attaches a long string of As to it, (the 'polyA' tail) and this is the basis of RNA sequencing. Reverse transcribe complimentary DNA (cDNA) to it. A third application is 'chip-seq' - trying to understand where on the DNA certain proteins might bind - link the proteins onto the DNA and crosslink the protein to the DNA in the cells. Chip-seq involves anti-bodies which pull out fragments of DNA which proteins are bound to. Methyl-seq determines where on the genome the DNA has been methylated (which proteins have been expressed in the cell). The methylation marks can be passed on from one cell cycle to another as cells divide - split DNA into two identical samples, and treat one differently (in a way which isolates Cs which bind to methyl groups).

## Reading for Week Two (hyperlinks)

- Explaining the 'sequencing-by-synthesis' methodology.
- Design and analysis of ChIP-seq experiments.
- An overview of NGS from a vendor.
- Information on the exact role of RNA in protein synthesis.
- Further information on the transcription process (DNA into RNA).

# Week 3: Computing Technology

We can broadly split the computing technologies into theory, systems and applications. Thinking computationally is essential. Don't forget: computers do exactly what you tell them to do! There are dozens, if not hundreds of programming languages: how we talk to computers. Specific defined terms with specific meanings - but writing good code is hard! Engineering means testing your code - robust, debugged. An algorithm describes what a computer can do: a step by step series of instructions on how to do something. Not necessarily used by computers: e.g. finding the maximum of a hill, or sorting a bunch of numbers. If your dataset is large, efficiency becomes key.

How do we get the sequencing data into memory? Most importantly, *we want to be able to find stuff* - e.g. 100 nucleoids. To understand how memory works, note that not only do we have the data itself, but also an address, termed 'pointers'. Numbers and characters are stored as 8 bits in a row - a byte - 0000 0000 - native computer code. Everything on a keyboard is represented as a single byte. Considering genomics specifically: DNA = {A,C,G,T} and A=00, C =01, G=10, T=11. When it comes to efficiency, think: how do they plan to deliver mail - when should the truck go back to the warehouse? It's important to understand that computer programs arent just black boxes - know what's going on under the hood. Of particular importance is the software behind 'alignment' - examples of RNA editing/differences across 'big' datasets. Typically 1 misalignment in 1 million: which yields 100s of errors in large genomics datasets. **Make sure your software handles all possible cases** - verify all your software if you think you have important results!

Computational biology software transforms raw data into information to guide discoveries: DNA sequencing data – A,C,G,Ts etc. There are a multitude of programs and pipelines. What's the difference from my genome than the reference genome? To find out: run it through a pipeline. Example – RNA-seq: used to measure difference between cell types, which genes are turned on, turn RNA into raw sequences, etc. Go from raw-reads to a list of genes and expression levels. 'Tuxedo tools': bowtie - alignment to human genome. tophat2 - spliced alignment. cufflinkks - transcript assembly and quantitation. Cuffdiff2 – differential expression between one set of data and another set of expressions. These programs are being superceded - bowtie2 - faster alignment, hisat - spliced alignment, ballgrown - differential expression - stringtie - transcript assembly and quantitation. What is surprising is that even though these are well defined, discrete tasks, different software make different alignments of the same reads - perhaps only 98-98.5% similarity? Even when aligning a read, the alignment may be to different places. Software is changing extremely quickly in this field (as is the data).

**In comparison to statistical/econometric estimation, where software will almost unilaterally produce identical results, this is not necessarily the case with computational biology toolboxes.**

## Reading for Week Three (hyperlinks)

- Further information the specifics of alignment.
- Details on hash based compression for storing DNA
- A guide to big data management with R
- An overview of RNASeq Analysis.

# Week 4: Why Care About Statistics?

Genomic datascience is based on three core disciplins of biology, computer science and statistics. Statistics is often overlooked in this triage: however, several existing studies are limited through their poor use of statistics which are often implemented poorly. This can even result in some lawsuits when results lead to clinical trials. There are two key reasons as to why statistical analysis can go wrong: a.) a lack of transparency and reproducibility of work done by others. This can also be further hampered by a lack of cooperation by leading authors. b.) The second key issue is the lack of statistical expertise and study design problems (e.g. batch problems - see below). Statistical expertise can help 'upstream' - before the study even begins (see power of tests - below). The central dogma of statistics:

**The central dogma of statistics involves inference on a population of interest from a carefully stratified sample. The key idea involves quantifying the variability of the sample.**

Data sharing plans are essential. All genomic datasets need four things:

1. Raw data: no processing, no computing, no deleting.
2. Tidy data: one variable per column, one observation per row, one table per 'kind' of variable, linking indicators per table.
3. Code Book: variable names, variable descriptions, study deisgn information.
4. Recipe to go from raw to tidy data with the code book: r/python code, inputs raw data and outputs tidy data with no parametres (avoid the temptation to not use a script).

An excellent example of a datasharing plan can be found here. This also introduces us to Github - a hugely valueable resource for code sharing and wikis. There are a multitude of resources available for learning statistics on the internet. One way to get a bit of help is via Cross Validated. A 'lonely bioinformatician' is a single statistician working in a lab without other statisticians (as opposed to a computational biology lab).

'Make big data as small as possible as quick as possible' to enable sharing and visualising. Sometimes statistical summary measures can be deceptive (hiding hidden patterns). One common, useful task is to 'plot replicates' - a very common plot where you compare the same sample being run through the technology twice. Log and other types of transforms are extremely useful. A common genomic example is an MA plot, which is an application of a Bland–Altman plot for visual representation of genomic data. The plot visualises the differences between measurements taken in two samples by transforming the data onto M (log ratio) and A (mean average) scales, then plotting these values. A common problem in genomics: a meaningless albeit visually impressive image of a network. Statistical graphs must not only look pretty, but convey scientific information to the reader.

Sample size and variability: with our best guest from a sample, we also get a guess about the variability. Typically, $N = \frac{\$youhave}{\$costpermeasurement}$. Even if means are different from each other, how confident can we be about that? This is measured with 'power': e.g., n=10, delta=5, sd=10:

```
power.t.test(n=10,delta=5,sd=10)
```

We can also 'back this out': how many observations will we need to have given a specific power, delta, sd, to detect a difference of specific magnitude? As you vary different paremtres (e.g. n, variance, etc), you get different powers: these calculations are therefore hypothetical based on what you think the effect size might be, and what power you might have. There are three types of variability in genomic measurement: phenotypic variability, measurement error and natural biological variation. Measurement error may decrease over time with technological advancements, but biological variation does not get elimiated by technology.

Statistical significance: are observed differences replicable? Are they real? A common metric that people use is the t-statistic:

$$t = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_y^2}{N} + \frac{S_X^2}{M}}} \tag{1.1}$$

This is the average of the Y minus the average of the X divided by some measure of variability: if they're very far apart in terms of variability units, we might think that they are statistically different. The most commonly used statistic is the p-value, calculated by 'scrambling' or permutating the relationship between the label and data of Y and X. Then, plot a histogram and see where the original value lands: calculate number of times the scrambled statistic is bigger than your observed value:

$$p = \frac{\#|S^{permutations}| \geq |S^{Obs}|}{\#permutations} \tag{1.2}$$

The p-value is the probability of observing a statistic that extreme if the null hypothesis is true. It is **not** the probability that the null or alternative is true, or a measure of statistical evidence.

However, this classical approach to p-values is not designed for multiple hypothesis tests at once. Note: if there is no differnce in what's going on, p-values are uniformly distributed - this means that 5% of the p-values will be less than 0.05. How do we correct for this? With different error rates:

- Standard p-value: 0.05×# tests = # false positives.
- False discovery rate $\leq$ 0.05× 550 = 27.5
- Family wise error rate controlled at 0.05: The probability of at least 1 false positive $\leq$ 0.05

However, it's important to report negative results and avoid p-value hacking - one way to do this is to state an analysis pla in advance of looking at your data, and stick to it. Confounding: a variable related to other variables which may look like there is a relationship when there isn't. One important confounder in genomics is the time at which the sample is taken (i.e. due to technological advancements). To overcome this, better consider your study and experimental design - through stratified sampling or randomziation. Other important things: balanced design (# treatments and controls), study should be replicated - technical replicates (measure how well technology works) and biology replicates (sample across people).

## Reading for Week Four (hyperlinks)

- An excellent discussion on p-values and power.
- A guide on 'gene set bagging' - to calculate probability of results will replicate in the future.
- A nice paper on the significance thresholds/MHT in GWAS studies.
- A related paper on power and significance testing in 'large-scale' genetic studies.
- Methods to overcome confounding.
- An overview of hypothesis testing (philosophical).