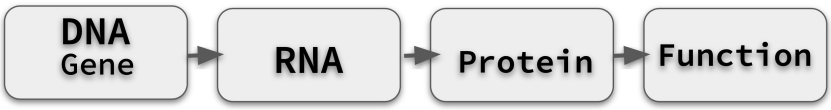
RNAge Supplementary Information

This notebook very briefly covers biological topics relevant to the RNAge project.

DNA and Genes

Deoxyribonucleic acid (DNA) in biology consists of four bases, ATGC, arranged in complementary sequences. The reference human genome consists of ~3.2 billion base pairs. However, not all of that long sequence necessarily *does* anything. There are special gene-encoding regions of the genome. Most of these regions encode for proteins, the functional molecules in a cell primarily comprised of amino acid chains.

The Central Dogma - how do genes become protein in a cell?



This is a simplified version of the central dogma which states that DNA, the linear *book* of information in each cell, is selectively **transcribed** into RNA, which is then **translated** into protein, the complex molecules which perform functions in biology.

Transcription is essentially copying a portion of DNA while translation converts the base-pair sequence into an amino acid sequence.

Though each cell in a complex organmism such as a human has the same DNA in all of its cells, the amounts of each gene that ultimately becomes protein is different.

Gene Expression

Differential expression of the genome is like reading from a book where the book is DNA in a cell. Each tissue has a differently annotated version of the same book; different tissues read different parts of the book. Thus, each tissue has a different fingerprint of which proteins it creates. In reality, this process is moderated by such factors as transcription factors/enhancers, the physical conformation of the DNA, histone modification, RNA interference, and other highly complex protein-protein interactions.

Given that gene expression underpins the cellular identity, researchers are interested in characterizing the gene expression of different cellular states. This approach has enabled breakthroughs in cancer research and disease pathology in general. Specifically, gene expression studies revealed the subtypes of breast cancer which are highly heterogenous. This breakthrough has enabled more personalized treatment of the molecular variant of breast cancer that a patient is facing.

There are different methods of accessing gene expression depending on where in the central dogma one focuses. Cancer gene expression sometimes be deduced from the deletions or modifications of the genome, but that wouldn't work for normal tissues. Researchers can measure levels of protein directly, but that can be extremely difficult as proteins are of all different shapes and sizes. In the end, measuring gene expression via the transcriptome balances robustness and cost.

Transcriptome

The transcriptome is the collection of all transcripts in a sample, which are just strands of RNA. Collecting and analyzing the transcripts in a sample is tantamount to analyzing the proteins in a cell because transcripts are assumed destined to become proteins. There are major exceptions to this rule nonetheless.

Since RNA is in the form of a nucleotide base sequence, knowing what gene a given transcript encodes is simply a matter of sequencing the transcript. Experimentally, mRNA, the stable version of transcripts, is extracted from a sample. These transcript are then sequenced using one of any Next-generation Sequencing platforms. The rawest form of the data consists 75-150 base-pair sequences. A common experiment may generate 50 million of these reads. These reads are then aligned to a genome, human in our case, to identify what gene each transcript fragment belongs to. The relative number of fragments aligning to a given gene is related to the abundance of that transcript, which as we know, is tantamount to knowing how much of that protein is being produced.

This process is called **RNA-sequencing** and will be the type of data used in this project.

RNA-sequencing and Normalization

The objective in a RNA-sequencing based gene expression study is to produce *counts* for each sample for each gene that reflects the intrinsic expression of that gene. Normalized counts are intended to be insensitive to experimental variance such as biological sample size, experimental bias, and sequencing depth.

TMM

The *trimmed mean of M-values* (TMM) normalization method, which can be read about in more detail here (https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25), normalizes the observed counts around the assumption that many genes are not differntially expressed across samples. The output counts are in the units of counts-per-million, or the number of aligned reads per million reads. This enables sample-sample comparison, but not gene-gene comparison as transcript length is not accounted for.

Gene naming conventions: HGNC and Ensembl

There are many well-characterized genes in the human genome, and there are even more uncharacterized gene-encoding regions. The fact that there can exist many transcript variants for each gene exacerbates the situation. Ensembl generates numeric names in the format of ENSG\<digits> for any genomic feature. These are not necessarily curated. They also genrate ENST\<digits> for transcript variants.

HGNC, however, generates curated gene symbols. A gene having a name does not guarantee that it has been characterized, but named genes are generally easier to remember and cross-reference. When possible, HGNC names are used, but the default is the automatic Ensembl name.