Genetics of Heart Development

Programming Analysis and Middle School Outreach

Justin Knight, Trevor Revesz Mentor: Jeff Kinne

THE CENTER FOR GENOMIC ADVOCACY AT INDIANA STATE UNIVERSITY

NIH National Institutes of Health

Sycamore Undergraduate Research Fellowship (SURF)

Indiana State University

Research Goals and Previous Work

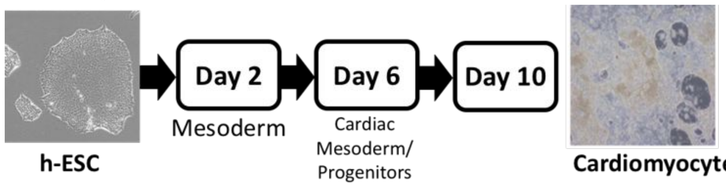
Goals

- ***Broad Goal*** Better understanding of **gene expression throughout heart development**, with potential implications for diagnosing and treating heart disease
- ***Ongoing Project*** Focus on **cardiomyocyte (heart muscle cell) development** beginning with mus musculus (mouse) embryonic stem cells and homo sapiens (human) pluripotent stem cells (not derived from embryos), knock down expression of suspected key genes to investigate its effects on cell development and expression of other genes
- ***This SURF Project*** Learn **R programming** and perform a different **analysis** to compare against results from the research team's SURE/BD4ISU 2018 results, develop materials and activities to **communicate research with middle school students**

Background, ZNF503

ZNF503, the mammalian ortholog of the Drosophila (fruit fly) noc gene, encodes a transcription factor that is a putative cardiac regulatory gene and is a gene of interest that is undergoing genetic loss-of-function experiments in both Drosophila and mammal experimental systems [1].

Experimental System



h-ESC (human embryonic stem cells) induced to become cardiomyocytes (heart muscle cells) [1]. Development was tracked for 10 days as h-ESC's are induced along the cardiomyocyte development pathway. RNA sequence data is extracted from the cells at key time points (days 0, 2, 6, 10) to study genetic differences at these time points. Image on right is a still of a video showing beating cardiomyocyte heart cells after 10 days of development. Experimental design and use of h-ESCs is from [1] and earlier work; Schwab Lab at ISU is successfully running the experiment with mouse ESCs and human induced pluripotent stem cells (not derived from embryos).

SURE/BD4ISU 2018 Analysis

A 2018 SURE/BD4ISU team analysed the data from [1]. Raw data were downloaded from the Gene Expression Omnibus (GSE69618). Before comparing wild type (normal) and ZNF503-knockdown samples, normalization of the expression values was performed using the edgeR and limma packages in R. The results were normalized log2 counts-per-million expression values. Differential expression analyses were performed using the same packages in R. **The normalization and expression analysis packages perform statistical and modeling calculations aimed to account for normal variation within samples and to give precise estimates for the probability that observed differences are not simply due to random variation.**

For one particular case (comparing normal and ZNF503-knockdown samples at day 6), the SURE/BD4ISU 2018 analysis identified 527 genes that were up-regulated in the normal sample by a factor of 2, and 285 genes that were down-regulated in the normal sample by a factor of 2 (each with a *p value* of at most 0.05, i.e. a probability of at most 5% that the observed difference is an artifact of random chance).

Our Analysis Approach

We ask whether the sophisticated packages used in previous analysis yield different results than a simpler approach. The raw data from GSE69618 contains raw counts of the number of RNA fragments of each gene present in the sample. For each time point (day 0, 2, 6, 10) and preparation (normal, ZNF503-knockdown) the data contains at least two replicates (meant to allow for accounting for natural random variation).

Our analysis consists of the following for each gene

1. **Average** the three replicates
2. Take the **logarithm** of the raw counts
   (not strictly necessary, but scales values similar to previous analysis)
3. **Subtract** day6 normal values from day6 ZNF503-knockdown values.

Genes where the difference in step 3 is at least 2 (or less than -2) are interpreted to result from a factor two up- (or down-) regulation of the gene in question in the normal samples.

Code, Data, and Analysis

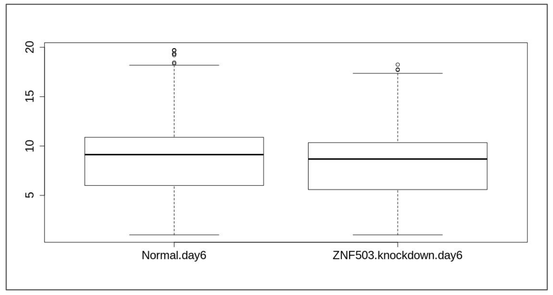
We use R programming to analyze the gene expression data. We load the data as follows.

```
data_full <- read.csv("GSE69618_data.csv")
data <- data[, c("geneID","Normal.day0","ZNF503.knockdown.day0",
                 "Normal.day6","ZNF503.knockdown.day6")]
data[,2:5] <- log2(data[,2:5])
```

geneID	*Normal.day0*	*ZNF503.knockdown.day0*	*Normal.day6*	*ZNF503.knockdown.day6*
NM_000014	2.0	1.0	8.4	4.9
NM_000015	4.9	5.0	2.0	2.6
NM_000016	11.5	11.6	11.6	10.3

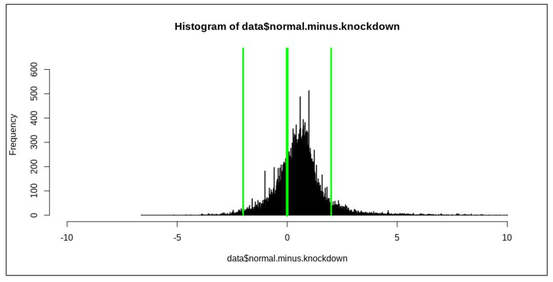
A box plot shows the 1st quartile, median, 3rd quartile.

```
boxplot(data[,c("Normal.day6", "ZNF503.knockdown.day6")])
```



We look at the difference in expression level between normal and knockdown.

```
data$normal.minus.knockdown <- data$Normal.day6-data$ZNF503.knockdown.day6
hist(data$normal.minus.knockdown, breaks=1000, xlim=c(-10,10))
```



There is a clear bias in the samples - there is either higher gene expression overall in normal samples, or this is an artifact of the sequencing from different samples.

We determined which genes were up-regulated in the normal samples by a factor of 2 by simply looking for a difference of 2.

```
surf_up_genes <- data[data$normal.minus.knockdown >= 2,]
```

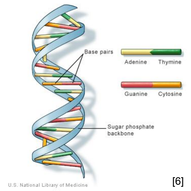
We used differences of 2, 3, or 4 to look for up-regulated genes, and -2 to look for down-regulated genes in the normal samples. We determined how many genes were identified with each cutoff and how many of the genes identified by SURE/BD4ISU 2018 appeared in our lists.

Cutoff	*# genes identified*	*# SURE/BD4ISU genes present*
2 (two-fold up-regulation in normal)	2730	476 out of 527 (90%)
3	1113	416 out of 527 (79%)
4	615	311 out of 527 (59%)
-2 (two-fold down-regulation in normal)	564	146 out of 285 (51%)

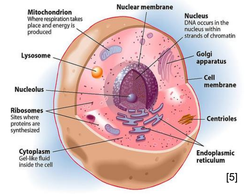
We conclude that our simpler approach may do a reasonable job of identifying up-regulated genes in the normal samples and clearly does not do well in identifying genes down-regulated in the normal samples. Further investigation is warranted.

Explaining Genetics to Middle Schoolers

We have developed a presentation and in-class activities to introduce middle school students to genetics in general and this project in particular. We **begin with basic biology terms and concepts** related to genetics and heart development - cell, parts of the cell, DNA, RNA, genes, transcription, translation, protein synthesis.

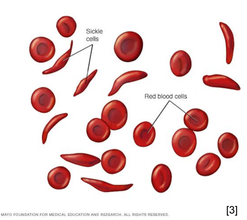


[6]

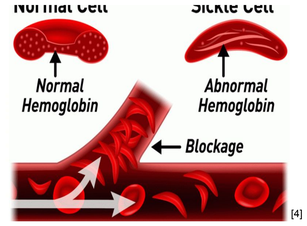


[5]

We use **sickle cell disease as a classic example of a genetic mutation** that causes disease. Sickle cell disease results from a single change in the genetic code for hemoglobin (a change in the *genotype*). The ***single nucleotide polymorphism*** (SNP, one change in the DNA/RNA sequence) results in a change of a single amino acid in the hemoglobin protein. The *phenotype* (resulting change in how things function) causes red blood cells to be misshapen, which can block blood flow throughout the body and cause pain.



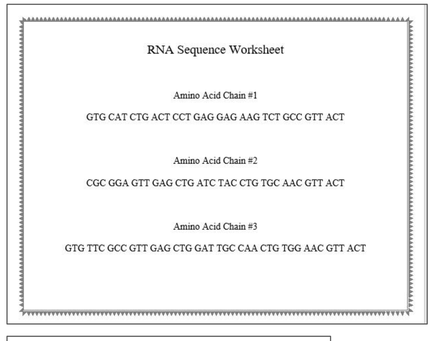
[3]



[4]

In Class Activity

In class activities reinforce the concepts as the students act out protein synthesis. We took an activity from [7] and adapted it. Proteins are synthesized by translating an RNA sequence into an amino acid sequence. Each 3 base pairs in the RNA sequence (called a *codon*) translates to 1 amino acid. We first translate RNA codons into English words so that the sequence results in a fun sentence. We then translate the RNA codons into amino acids. **We use the actual RNA sequence for the hemoglobin gene**, with some groups having the normal sequence and some groups having the sequence that results in sickle cell disease.



RNA Sequence Worksheet

Amino Acid Chain #1

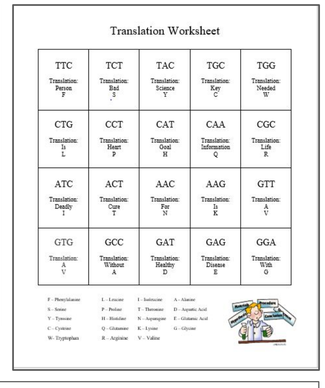
GTG CAT CTG ACT CCT GAG GAG AAG TCT GCC GTT ACT

Amino Acid Chain #2

CGC GGA GTT GAG CTG ATC TAC CTG TGC AAC GTT ACT

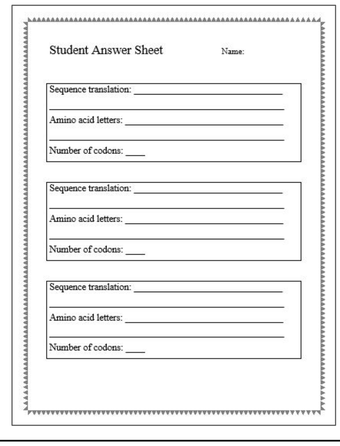
Amino Acid Chain #3

GTG TTC GCC GTT GAG CTG GAT TGC CAA CTG TGG AAC GTT ACT



Translation Worksheet

TTC	TCT	TAC	TGC	TGG
CTG	CCT	CAT	CAA	CGC
ATC	ACT	AAC	AAG	GTT
GTG	GCC	GAT	GAG	GGA



Student Answer Sheet Name:

Sequence translation: _____

Amino acid letters: _____

Number of codons: ____

Sequence translation: _____

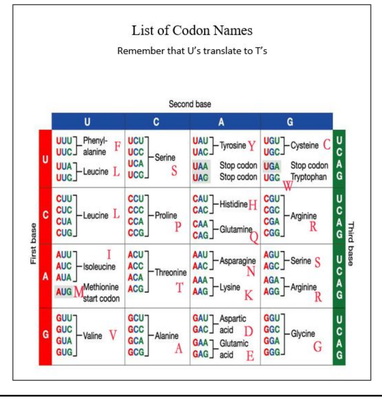
Amino acid letters: _____

Number of codons: ____

Sequence translation: _____

Amino acid letters: _____

Number of codons: ____



List of Codon Names

Remember that U's translate to T's

Second base

First base	U	C	A	G	Third base
U	UUU Phenylalanine (F)	UCU Serine (S)	UAU Tyrosine (Y)	UGU Cysteine (C)	U
	UUC	UCC	UAC	UGC	C
	UUA Leucine (L)	UCA	UAA Stop codon	UGA Stop codon / Tryptophan (W)	A
	UUG	UCG	UAG Stop codon	UGG	G
C	CUU Leucine (L)	CCU Proline (P)	CAU Histidine (H)	CGU Arginine (R)	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Glutamine (Q)	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Isoleucine (I)	ACU Threonine (T)	AAU Asparagine (N)	AGU Serine (S)	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lysine (K)	AGA Arginine (R)	A
	AUG Methionine / start codon (M)	ACG	AAG	AGG	G
G	GUU Valine (V)	GCU Alanine (A)	GAU Aspartic acid (D)	GGU Glycine (G)	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glutamic acid (E)	GGA	A
	GUG	GCG	GAG	GGG	G

Acknowledgments

This project was part of an ongoing collaborative research project led by ISU faculty members Drs. Shaad Ahmad, Jeff Kinne, Kristopher Schwab, and Catherine Steding. The 2018 SURE and BD4ISU team included ISU students Laura Cochran, Joseph Dalloul, Hayden Fell, Garett Oxford, Josh Soto, Dwayne Tally, Michaela Ward, and Andrew Williamson; the following posters were produced - "A Gene Expression Atlas of Embryonic Cardiac Genes Generated by Developmental Stage-Specific Analysis" and "Differential Gene Expression Analysis of ZNF503 Loss-of-Function in Human Cardiogenesis".

The project has been funded by the Center for Student Research and Creativity (CSRC), NIH grant 1R25MD011712-01 (BD4ISU), University Research Council (URC), Department of Mathematics and Computer Science, The Center for Genomic Advocacy (TCGA), and Department of Biology.

References

[1] Busser BW, Lin Y, Yang Y, Zhu J et al. An Orthologous Epigenetic Gene Expression Signature Derived from Differentiating Embryonic Stem Cells Identifies Regulators of Cardiogenesis. *PLoS One* 2015;10(10):e0141066.
[2] Marianne Williams, "What is the purpose of DNA transcription?", accessed on 14 Nov 2018 at https://www.quora.com/What-is-the-purpose-of-DNA-transcription
[3] Mayo Clinic, "Sickle Cell Anemia", accessed on 27 Nov 2018 at https://www.mayoclinic.org/diseases-conditions/sickle-cell-anemia/symptoms-causes/syc-20355876
[4] The New Journal and Guide, "Va. Takes Lead In Sickle Cell Disease Relief", accessed on 27 Nov 2018 at http://thenewjournalandguide.com/va-takes-lead-in-sickle-cell-disease-relief/
[5] 625 Points, "Cell Structure: Notes and Sample Answers", accessed on 12 Nov 2018 at http://www.625points.com/2016/08/leaving-cert-biology-cell-structure.html
[6] Genetics Home Reference, "What is DNA?", accessed on 10 Nov 2018 at https://ghr.nlm.nih.gov/primer/basics/dna
[7] Teachers Pay Teachers, "Thanksgiving DNA, RNA & Protein Synthesis Game/Activities", accessed on 4 Nov 2018 at https://www.teacherspayteachers.com/Product/Thanksgiving-DNA-RNA-Protein-Synthesis-GameActivities-2857535