Computational Biology Camp

Cheat Sheet

[Meghna Kannan - 6/17/24]

MIRcore

Lecture Notes - Genomics

- Precision Medicine: study of application of individual variability in biological traits
  - Same treatment ⇨ different effect (trial & error)
  - Precision medicine is when there is no trial & error
- Gene: hereditary unit that codes for a functional molecule
- Transcription and translation: DNA & RNA:
  - DNA always lives in nucleus, small sequences of this DNA transcribed into RNA, the RNA goes to the ribosome and codes for protein using translation
- Genome: all of an organism's DNA ⇨ protein coding, regulatory sequences, other RNAs
  - Genomics: study of genomes
  - Transcriptome: RNA level
  - Protein Level

- Cost of DNA sequencing has reduced over years
- Genome Sequencing steps:
  - Break Genome
  - Order clones
  - Break clones
  - Generate and assemble sequence clones
  - Assemble sequences of overlapping clones ⇨ reference sequence
- Genome sequencing can help diagnose a disorder ⇨ shows how tailored treatment is possible
- GINA: Prohibits discrimination in health coverage or employment based upon genetic tests
- NIH is the biggest funder for research for genomics
- Colonialism in genome sequencing: because only certain ethnic groups were being sequenced, the medicine may not work on everyone (equitable access). Can be used in discriminating and racist ways.

- Current Medicine: Trial and Error
- Genomic makes precision medicine possible
- Gene sequencing is cheaper than ever

Lecture Notes - Genetics at a Molecular Scale

- DNA: deoxyribonucleic acid; polymer made of monomers A, T, G, & C
  - Adenine, Guanine, Cytosine, & Thymine
  - 3 components:
    - Ribose
    - Base
    - Phosphate - always carries a negative charge
- A-T Dimer Formation: reaction between 5’ and 3’, we read 5’ to 3’
- Bases connect between sugar and phosphate
- Each base bring a -1 charge to the strand
- Partial charges can attract ⇨ hydrogen bonding is a term for an attraction between a hydrogen and another atom (happens with the bases)
- A & T always hydrogen bond, G & C always hydrogen bond

- Antiparallel: we read 5’ to 3’, opposite strand reads 3’ to 5’
- Nucleosome: histone cores (histone proteins) are positively charged and DNA wraps around it, becomes and chromosome eventually
- Humans have 23 pairs of chromosomes: 22 autosome pairs, 1 sex chromosome pair
- Amount of DNA for a specific gene varies for the gene
- Chromosomes have short (p) arm & long (q) arm
- miRNAs: scientists have identified location where they occur; multifactorial are complex because the cause can’t just easily be pointed out

Lecture Notes - Research

- Misconceptions about genes and diseases:
  - Since genes come from parents, genes are faet that we cannot do much about
  - If genes are related to a disease, they are bad
- Gene Facts:
  - We cannot easily change gene sequences, but we can change gene expressions
  - All genes have a certain function
  - Genes work together with other genes (networks)

- Systems Biology: The whole is greater than the sum of its parts
- Diseases: conditions that occur when our body fails to maintain homeostasis
- Genes in DNA level and Diseases:
    - Almost all genetic tests check DNA sequence

Lecture Notes - Central Dogma

- DNA ⇨ RNA ⇨ Protein
- Enzymes:
  - Polymerase: synthesize long chains of nucleic acid
  - Nuclease: cleave phosphodiester bonds between bases
  - Ligase: joining two large molecules by forming a new chemical bond
  - Helicase: separates two hybridized nucleic acids
- Most genes are transcribed from **one** of the two strands
- Plus is DNA with 5’ end in the P arm of the chromosome (nothing to do with charge)
- RNA:
  - Coding RNA: messenger RNA, eventually codes for proteins
  - Non-coding: transfer RNA

- Splicing: two general; regions:
  - exons: end up in final product
  - intons: fold into itself and is discarded (go **into** the trash)
  - still have start and end codons
  - lower case: introns; upper case: exons
- DNA can make many different proteins through alternative splicing:
  - Certain exons aren’t included for specific proteins, based on what is defined as an exon; same long sentence, different smaller sentences and phrases.
- DNA has same sequences in different cells, RNA has different number of different sequences
- different relevance of genes for different functions in each cell

Lecture Notes - Statistics

- Null hypothesis: In general, is almost like the opposite of what you are trying to prove
- P-value:
  - Assume null hypothesis is true
  - Ask what's the probability of getting those result if the null hypothesis is true; probability due to chance?
  - If small P-value ⇨ reject null
- Three keys numbers
  - Mean (measure of center)
  - Standard deviation (measure of spread)
  - Sample size (gives statistical power)
- If the p-value's low ⇨ the null must go!
- If the p-value's high ⇨ the null is your guy!!

- Logarithms: opposite of exponents
  - Can be used in the richter magnitude scale (used to determine magnitude of earthquakes)
  - Fold change: b/a = b gene expression mean / a gene expression mean
- T-test: to see if there is a significant difference between the means of two groups

Disease Research

Useful websites: NIH, Medline Plus, Healthfinder, NCBI

OMIM =

[how to use OMIM]

Phenotype, molecular basis known

% Phenotype, molecular basis unknown

+ Gene and Phenotype combined

* Gene description

My selected disease: Glioblastoma

Symptoms:

Diagnostics

Treatment

Causes:

Genes

Social Impact

Gene Expression/Microarrays

- Traditional microarrays correlate light intensity to expression; used in a comparative sense, relative difference is important
- Picking a good data set:
  - at least 20-30 samples
  - experimental/control grouped ***and*** distributed
  - clear definition
  - all variables
  - Filter
- Greyscale microarray
  - Diameter is correlated to fluorescence
  - Fluorescence is correlated
- Use GEO2R for biomarkers

- Up Regulation ⇨ gene is being expressed more in certain cell (more mRNA product showed up in microarray)
  - Down regulation is the opposite
  - Any positive log(fc) value means it is up regulated
- Volcano plot tells both the significance ***and*** change in expression for genes in the dataset (click explore and download for individual points)
- UMAP plot: look at all colors grouped together

Lecture Notes - Groups of Genes

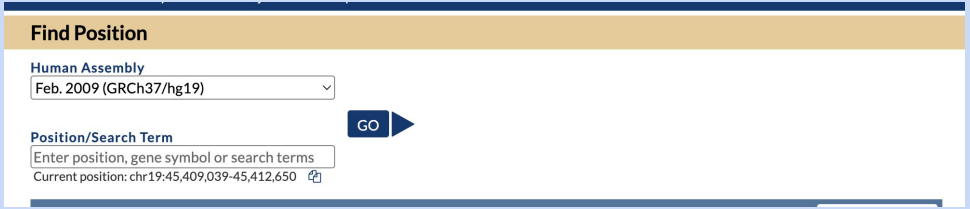
- Ways of selecting genes based on statistical outputs
  - T-test p-value < 0.05
  - Top 250 most significantly expressed genes
  - Up-regulated genes ($p < 0.01$) in smoker cancer patients
  - Absolute of log(FC) > 1
- Ontology: field of study focused on classification (ex. Location, organism)
- GO ⇨ gene ontology (classification of genes)
  - Biological processes: set of molecular events defined with start/end
  - Cellular components: part of a cell or its extracellular environment
  - Molecular function: activities of a gene product at molecular level

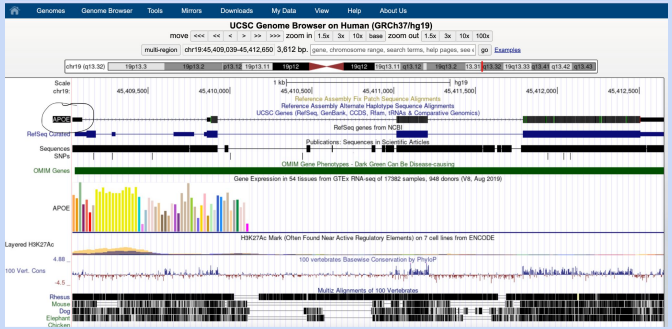
Lecture Notes - Gene Regulation

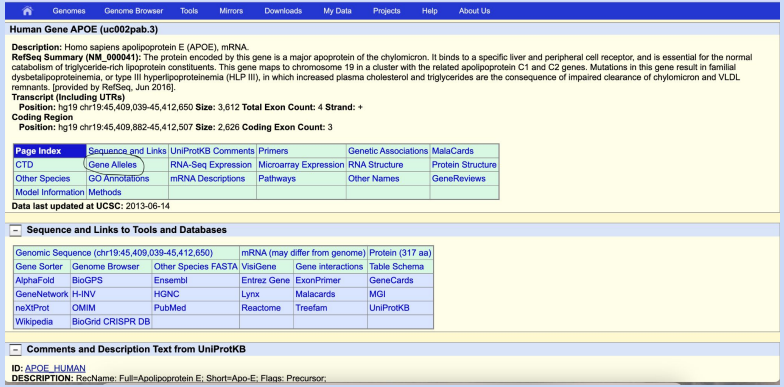
- Gene regulation: promoting or preventing gene expression
  - Can happen at the DNA level
    - methylation: less likely to be transcribed, DNA level
  - Can happen at the RNA level
    - microRNAs can prevent gene expression
    - small interfering RNAs (SIRNAs)
  - Destruction of one gene can impact other genes too ⇨ larger scale affect ⇨ homeostasis
- Signaling pathways
  - bodies are working through signals
  - signaling pathways show how the signals pass down through various gene products (proteins or RNAs)
  - two main types: activation and suppression

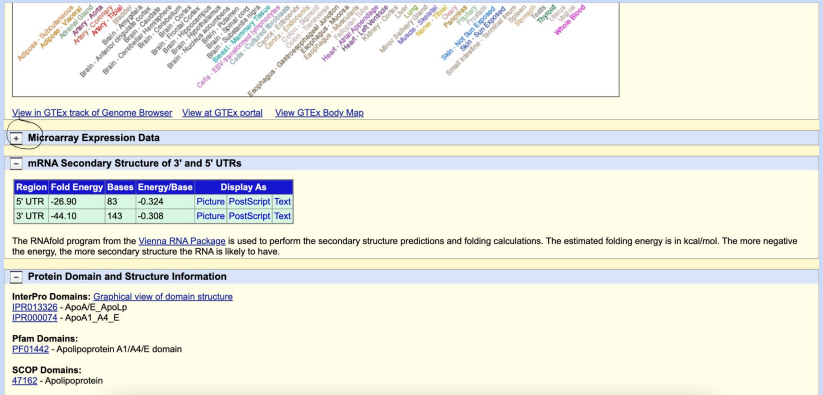
UCSC Genome Browser

Definitions: introns/exons, CDS, UTR









GEO + GEO2R

Definitions: p-value, null hypothesis, LogFC, up/downregulation

What to look for in a dataset

GEO2R steps

GO + KEGG

[GO Steps/Screenshots]

[KEGG Steps/Screenshots]

Some interesting facts about the small cell lung cancer pathway:

- All of the genes connected to either ECM or CDK4/6
- All of the genes had really strong connections
- Most pathways led to ECM and CDK

String-db

Definitions: enrichment/depletion

[Steps/Screenshots]

1.) Search protein and organism
2.) Press continue
3.) Can look at specific interactions
4.) For multiple proteins ⇨ choose multiple proteins instead of protein by name

How to Read Research Papers

Introduction

Methods

Results

Discussion

Conclusion

Acknowledgements

References

Figures

Abstract