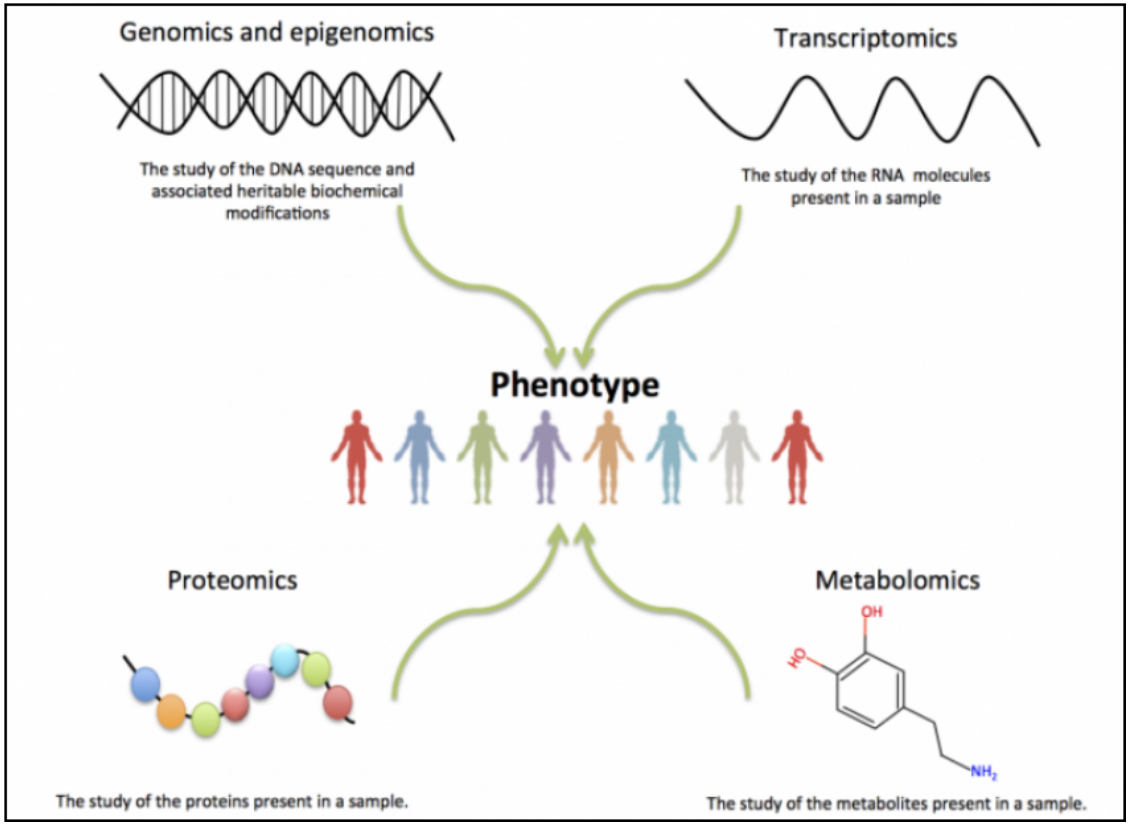
CS123A Bioinformatics Module 4 – Week 13 – Presentation1

Leonard Wesley

Computer Science Dept

San Jose State Univ



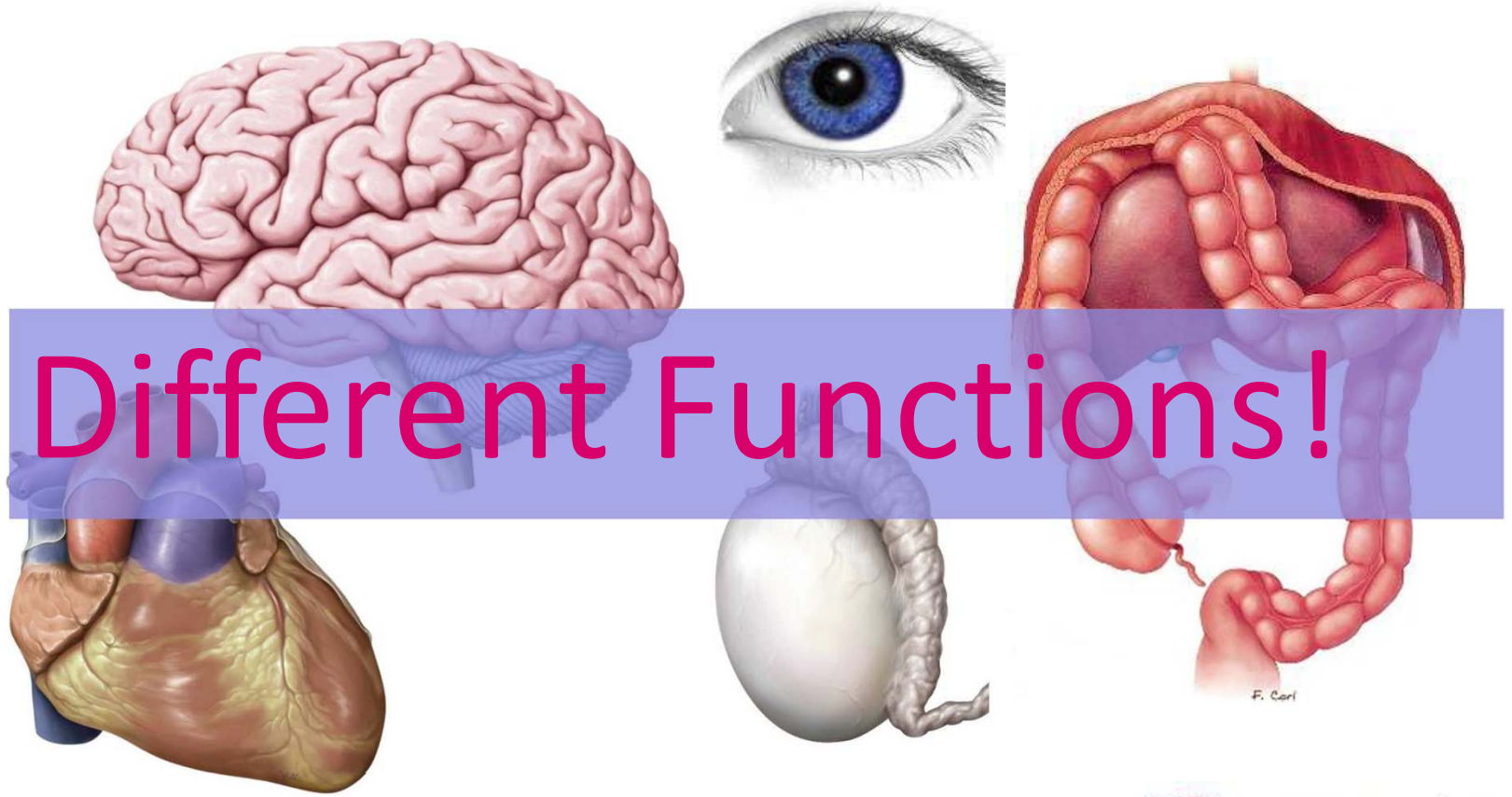
Agenda

- Introduction To Functional Genomics
  - Chap 14 in textbook.

What accounts for the difference in phenotype?

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE | CAMBRIDGE INSTITUTE

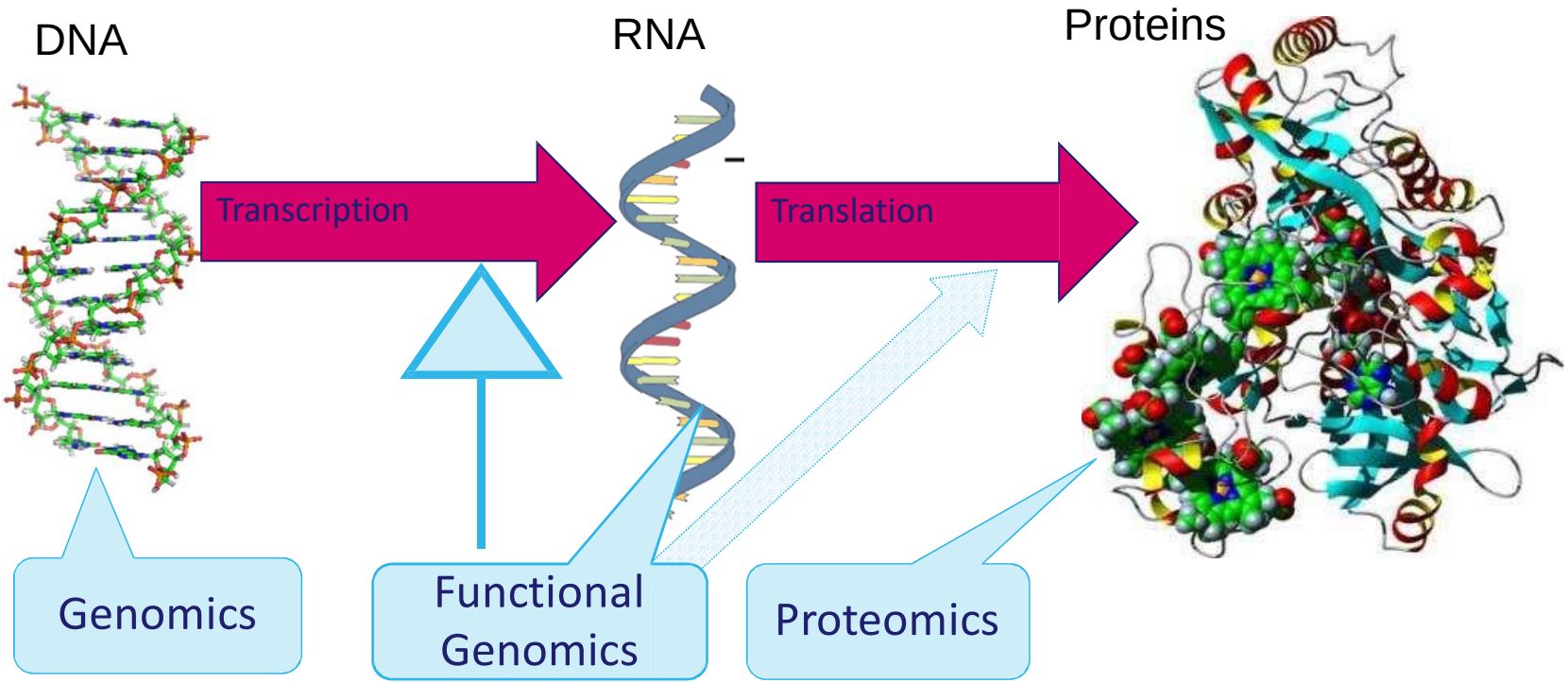
What accounts for the difference in phenotype?



Different Functions!

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

The Central Dogma of Molecular Biology



Definition Of Functional Genomics

- The genome-wide study of the function of DNA (including genes and nongenic elements) as well as the nucleic acid and protein products encoded by DNA.
  - Functional genomics may be applied to the complete collection of DNA (the genome), RNA (the transcriptome), or protein (the proteome) of an organism.
  - The assessment of RNA transcripts that are expressed at various times of development or various body regions constitutes an example of functional genomics.
  - Functional genomics often involves the perturbation of gene function to investigate the consequence on the function of other genes in a genome.

A great challenge of functional genomics is to understand the relationship between genotype and phenotype.

Functional genomics involves implementing experimental, computational, and analytic strategies to elucidate the function of DNA and chromosomes in relation to phenotype at the levels of the cell, the tissue, and the organism.

Reverse and Forward Genetics

- Reverse genetics asks “What is the phenotype of this mutation?”
  - Reverse genetics approaches attempt to generate null alleles as a primary strategy (and conditional alleles in many cases). *A null allele is a nonfunctional allele (a variant of a gene) caused by a genetic mutation.*

- Forward genetics asks “What mutants have this particular phenotype?”
  - Forward genetics strategies such as chemical mutagenesis are “blind” in that multiple mutant alleles are generated that affect a phenotype.

Reverse and Forward Proteomics

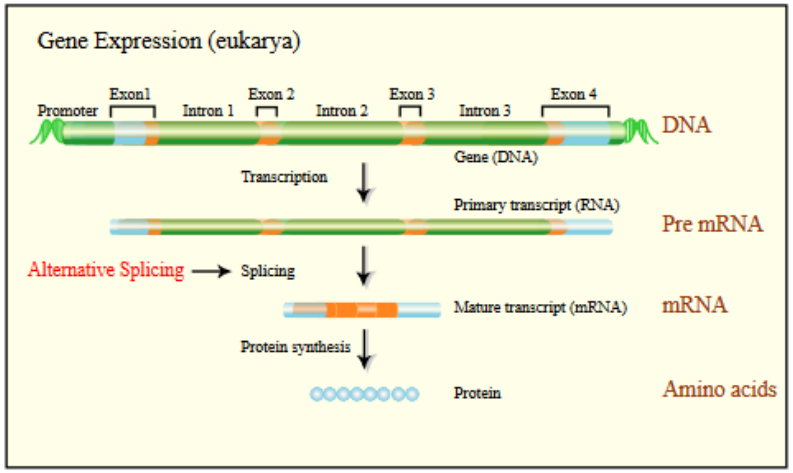
- Forward proteomics approaches correspond to the classical approach to protein characterization
  - A biological system is selected, such as human cells from individuals with or without a disease.
  - Proteins are compared by techniques such as mass spectrometry, differentially regulated proteins are identified, and from this the function of these proteins and their possible roles in the disease state may be inferred and further studied.

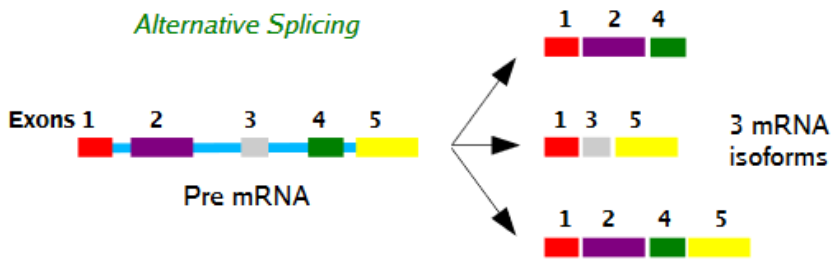
- Reverse proteomics start with genomic sequence from which genes, RNA transcripts, and protein products can be inferred.
  - Complementary DNA (cDNA) clones can be obtained and expressed in a variety of systems so that their function may be assessed in assays for protein–protein interactions or other behaviors (cellular phenotypes).

Some Definitions

- Genome: An organism's complete set of DNA, including all of its coding and non-coding genes.

- Transcriptome: All mRNA present in a cell at a particular state.





Different isoforms -> different function, i.e., different proteins translated.

Figure by MIT OCW.

Coding and non-Coding RNA Categories

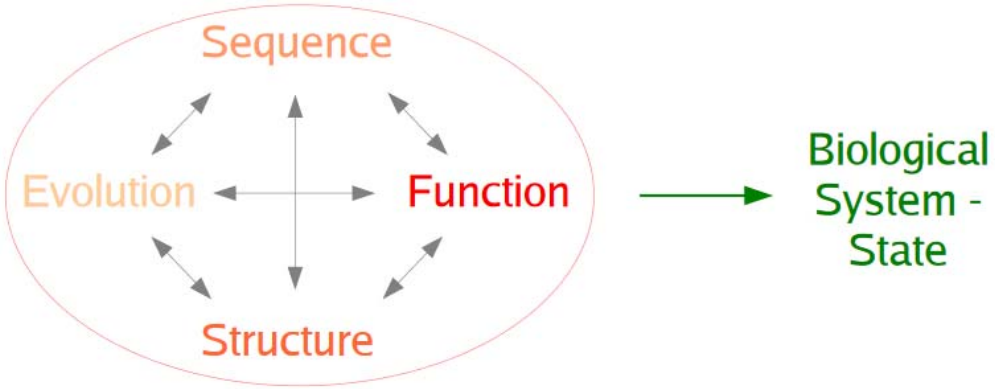
- Messenger RNA = protein coding transcripts, typically high degradation rate
- Transfer RNA = transfer AA to polypeptide chain during translation
- Ribosomal RNA = primary (structural) constituent of ribosomes
- Small nuclear RNA = RNA splicing, telomere maintenance, form snRNProteins (“snurps” small nuclear ribonucleoproteins)
- Small nucleolar RNA = chemical modification (e.g. methylation) of rRNA
- Guide RNA = RNA editing in protozoa
- Micro RNA = RNA interference at post/pre-transcription

cDNA and Expressed Sequence Tags

- Recall, transcriptome = all mRNA present in a cell at a particular state, organism-space-time specific.

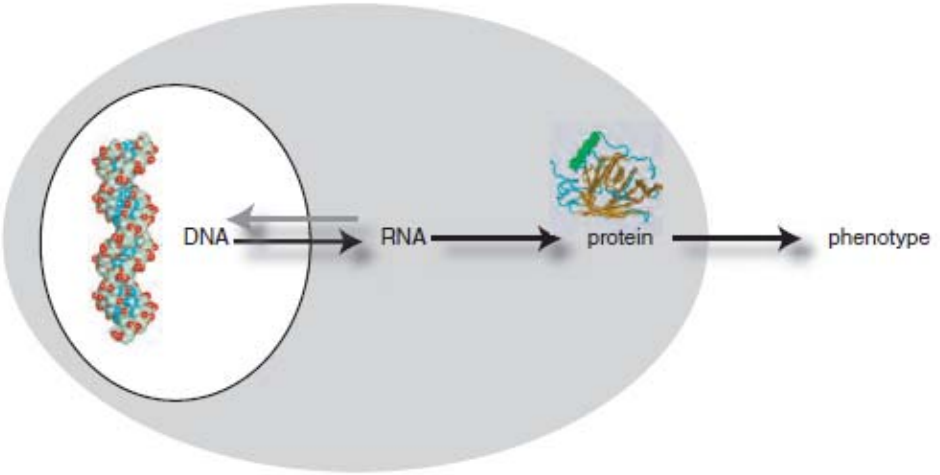
- Identification / characterization
  - Genomic libraries: DNA fragments of (near) total genome @ specific state
  - cDNA libraries: mRNA fragments (no intron) -> cDNA fragments -> sequence -> expressed sequence tags (EST's), GenBank ID#
  - 1 gene “covered” by >1 EST's. Eg. human genome >4M EST's, ~25K genes
  - Screen EST's -> EST's associated with a particular gene form a Unigene cluster
  - Differential comparison between cDNA libraries: Binary analysis (present/absent). H_0: # of sequences for a given gene X is the same in two libraries. Prob test: Fisher exact. Limitations: sequencing error + depth, tissue of origin contamination, and library construction bias.

Different Scales Of Function For A Biomolecule X



- Chemical/physical (microscopic scale): binds another molecule, catalyzes a molecular reaction, etc.
- Biological (macroscopic scale): leads to a phenomenologic / phenotypic transformation
- All scales in between the above (mesoscopic)
- X may have >1 function, across / within these scales
- A general / naïve test for function: Perturb X in native system and observe what happens at all scales

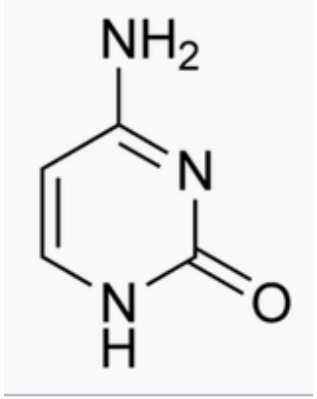
Functional Genomics Approaches To High-Throughput Protein Analysis



	DNA	RNA	protein
Natural variation --across development --across body regions --across species, strains	SNPs; epigenomics	transcriptome profiling (RNA-seq)	protein localization; protein-protein interactions; pathways
Functional disruptions --experimental	knockout collections transgenic animals	RNAi; siRNA	chemical modification
--in nature	Williams syndrome Down syndrome cancer chromosomal changes	nonsense-mediated RNA decay	myasthenia gravis

Example: Find The Function Of A Gene

- Eg. mutation (frameshift, mis-sense / non-synonymous) of methyl-CpG-binding protein 2 (MECP2, Xq28) -> Rett syndrome
  - Rett Syndrome is a rare X-dominant non-inherited genetic postnatal neurological development disorder that occurs almost exclusively in girls and leads to severe impairments, affecting nearly every aspect of the child's life: their ability to speak, walk, eat, and even breathe easily. The hallmark of Rett syndrome is near constant repetitive hand movements while awake.
- Phenotype incl. autism, dystonia, short, etc. Typ. fatal in males (major encephalopathy).



Cytosine

Example: Find The Function Of A Gene *(cont.)*

- MECP2 chemical function: binds methylated DNA -> repress transcription from methylated gene promoters
- MECP2 biological function: embryonic development
- Another example mutation (truncating frameshift, mis-sense) of cyclin dependent kinase-like 5 (CDKL5, Xp22) leads to almost similar phenotype. CDKL5 chemical functions: ATP binding, protein serine / threonine kinase activity, nucleotide binding.

Typical Questions In Functional Genomics

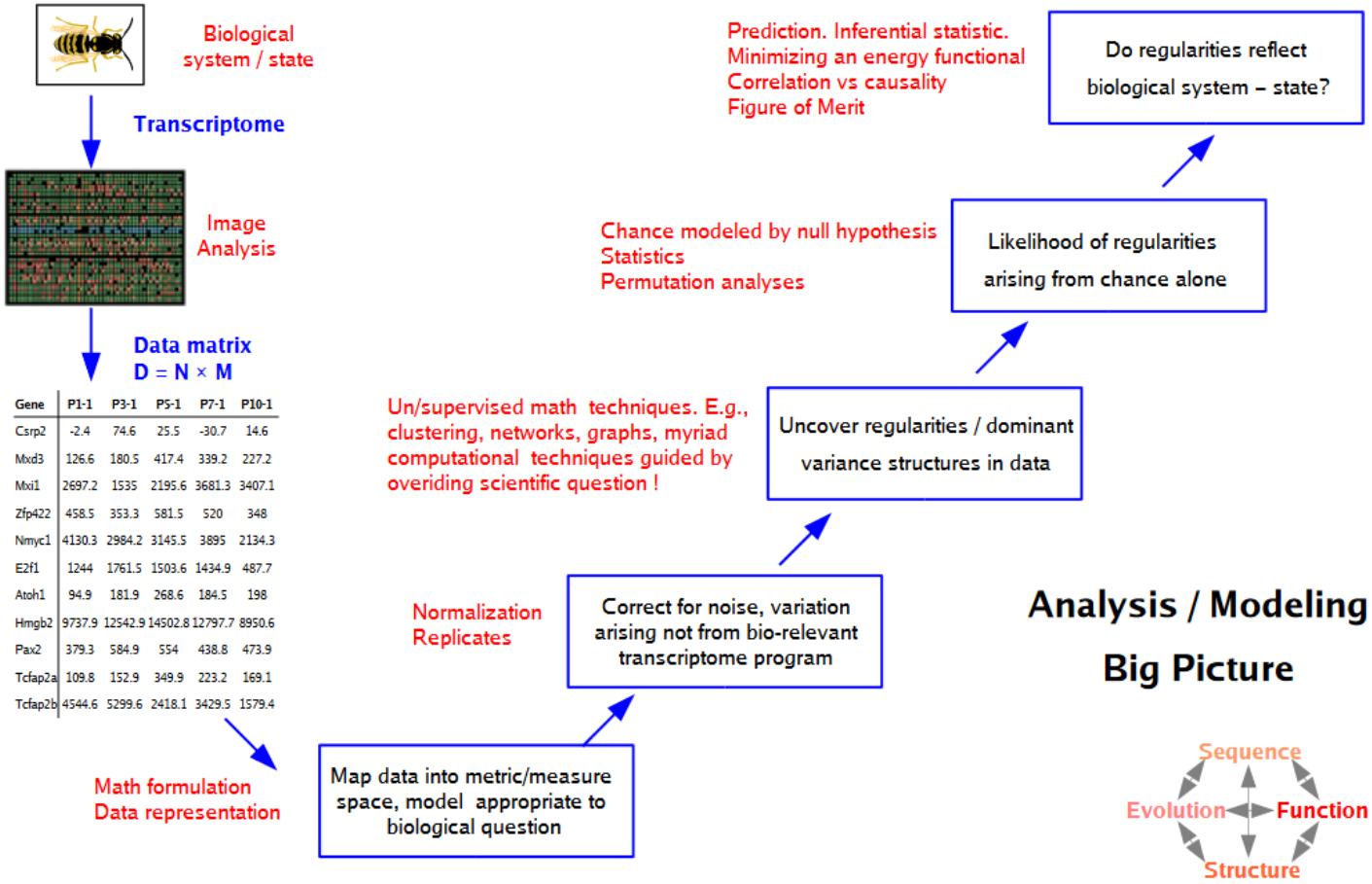
- What function does a given molecule X have in a specific biological system - state?

- Which molecules such as non-genic/chemical-molecules/genes/proteins (their interactions) “associate” with / “underwrite” a given biological system - state?

- What is in our genome (e.g., eukaryotes)
  - Genes (~1.5% genome. Eg. protein coding exons),
  - Gene-related DNA(~36% genome. Eg. non-coding introns – eukarya, pseudogenes),
  - Intergenic DNA (~62.5% genome. Eg. microsatellites, genome-wide repeats).
  - Coding = transmission into mRNA.
  - Genome-wide repeats. E.g., transposons, long/short interspersed nuclear elements

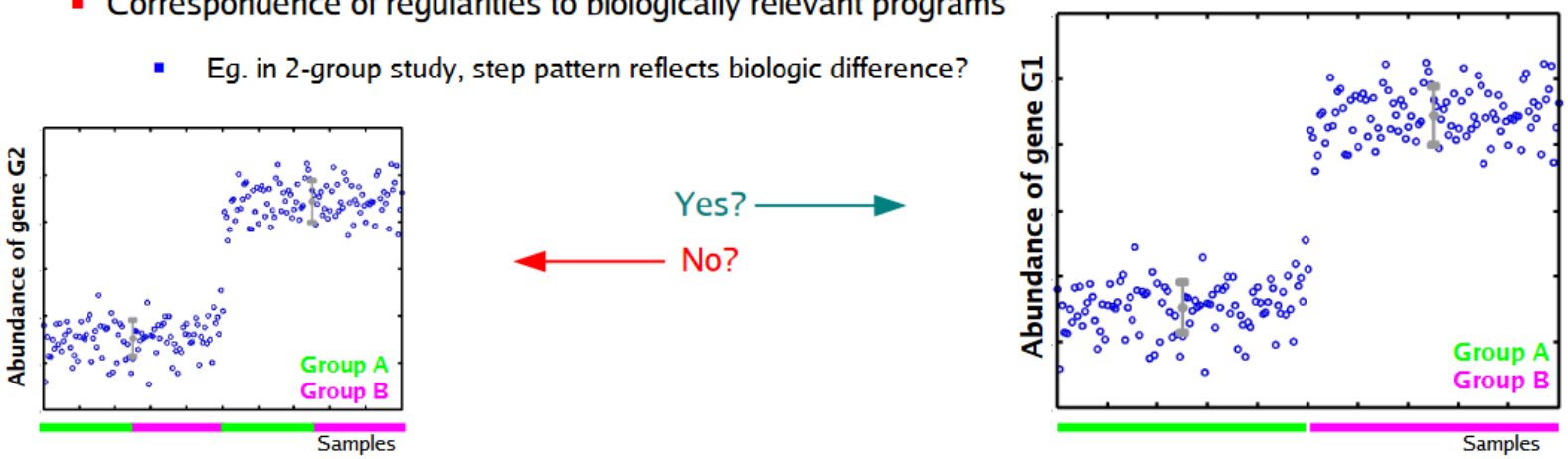
http://www.ncbi.nlm.nih.gov/Tools/

Functional Genomics Data Flow Analysis



Regularities In Data

- *Regularities* refer to dominant variance structures or coherent geometric structures intrinsic to data with respect to a particular measure / metric space.
- An observed pattern may be regarded a regularity if the pattern can be correlated to *a priori* scientific knowledge. Caveat: bias (*supervised* analysis to additionally test for bias)
  - Eg. in a 2-group comparison study, k of N genes were found to be differentially expressed between groups. "Step" function pattern for each gene is a regularity.
- Statistical likelihood of obtaining regularities given the data distribution
  - *A priori* knowledge/assumptions of underlying distributions to form relevant null hypotheses. Internal correlations and structural assumptions to reduce theoretical degree of freedom – modifying null hypothesis
  - Multiple testing. Bonferroni-type corrections: (type 1 error / false +) $\alpha \rightarrow \alpha/(\#$ of times test applied)
- Correspondence of regularities to biologically relevant programs
  - Eg. in 2-group study, step pattern reflects biologic difference?



Functional Genomics Exercise: Finding The Phenotype Of A SNP

Finding The Phenotype Of A SNP Scenario

- You are working with the sequence variation **rs2068824**.
  - This is a *Single Nucleotide Polymorphism* (*SNP*) located in the MMEL1 gene in humans. The MMEL1 gene codes for a metalloprotease that cleaves polypeptides preferentially between hydrophobic residues.
  - This sequence variation has turned up in several of your samples in patients with digestive troubles. We will use *Ensembl* to find out if any phenotypes or diseases are known to be associated with this SNP.

- Lets find out about the MMEL1 gene.
- Go to 'NCBI Gene' and enter MMEL1 in the search window and click SEARCH.
- Click the first entry for Homo sapiens/Humans.

Finding The Phenotype Of A SNP Scenario *(cont-1)*

https://www.ncbi.nlm.nih.gov/snp/

- Go to the NCBI SNP DB
- Enter **rs2068824** into the search window and click SEARCH
  - Variant type = SNV means each SNP is present to some appreciable degree within a population (e.g. > 1%).
  - GMAF/MAF = Global Minor allele frequency. For Example, C=0.0545/210 (ALSPAC) means 5.45% of ~2,400 subjects or 210 total chromosomes were observed to have the T>C SNP.
- Click on the GeneView link on the "Gene: NAV1(GeneView)" line.
  - Hover over the "rs2068824" (in red) symbol. Then when the pop-up window appears, click on the SNP Summary link.
  - NOTE: "Variation Type", "Gene: Consequence", … etc.
  - Click on "Frequence" tab on left to get info RE frequence of occurrence in different populations.

Finding The Phenotype Of A SNP Scenario *(cont-2)*

Lets use Ensembl-EBI DB: Hosted in the UK, Created in 1999, provides a genome browser that acts as a single point of access to annotated genomes for mainly vertebrate species.

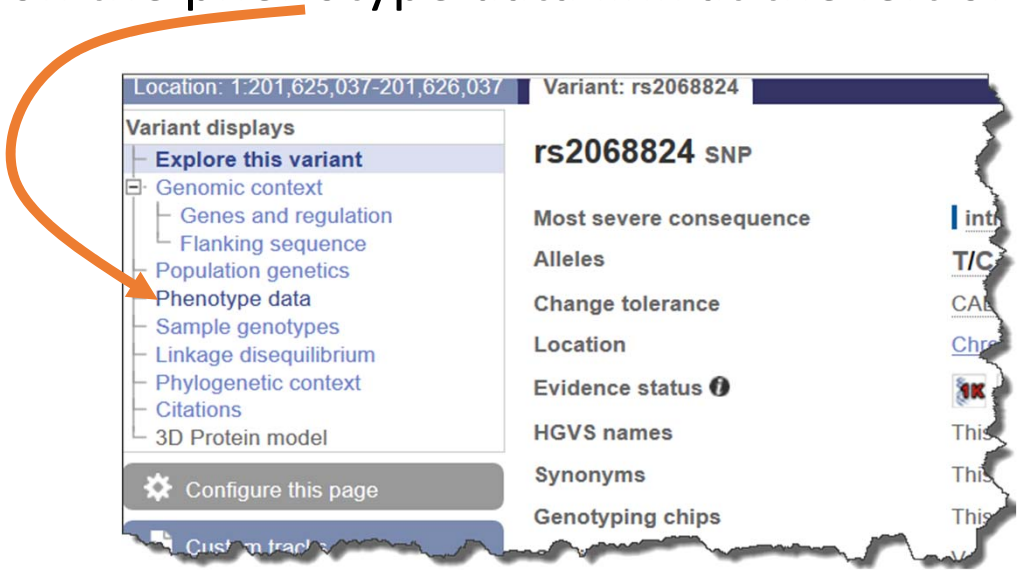
The Ensembl DB includes information on protein domains, genetic variation, homology, syntenic regions and regulatory elements. Coupled with analyses such as whole genome alignments and effects of sequence variation on protein, this powerful tool aims to describe a gene or genomic region in detail.

http://www.ensembl.org/

- Go to the Ensembl homepage
- Select Human species and enter 'rs2068824' in the search box.
  - NOTE: Esembl has some of the same and related info as the NCBI SNP DB
- Click on the “rs2068824 (Human Variant)” link

Finding The Phenotype Of A SNP Scenario *(cont-3)*

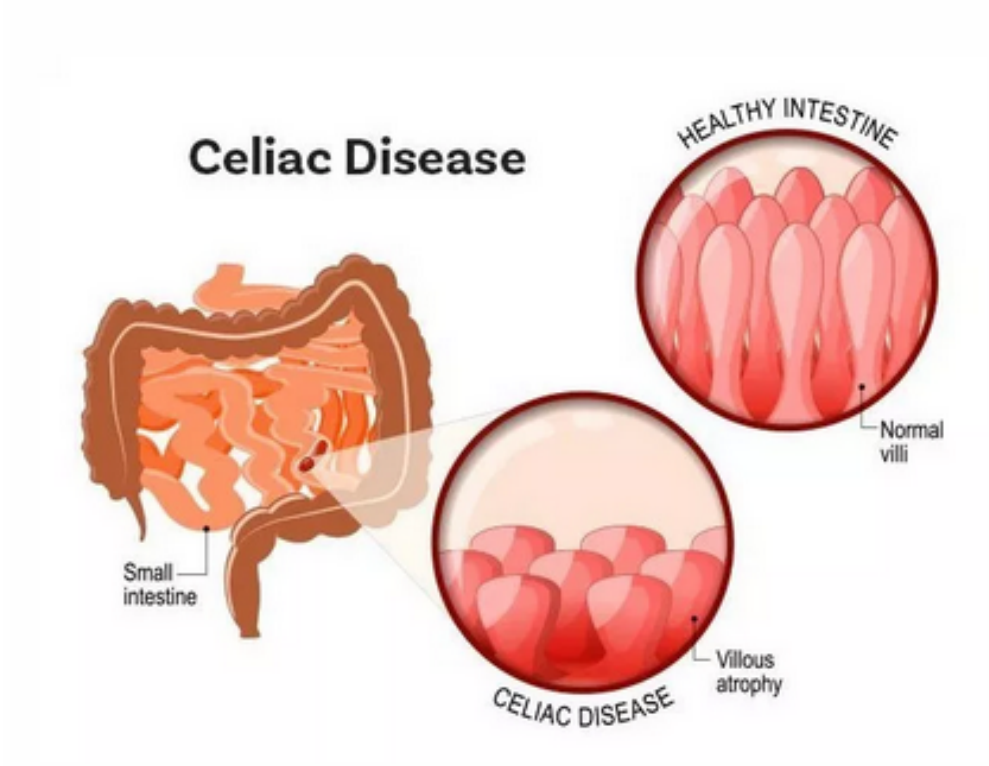
- Click on the phenotype data link at the left of the “Variant rs2068824” tab



- Scroll down to the “Phenotype Data” section.
  - The associated disease is Celiac disease

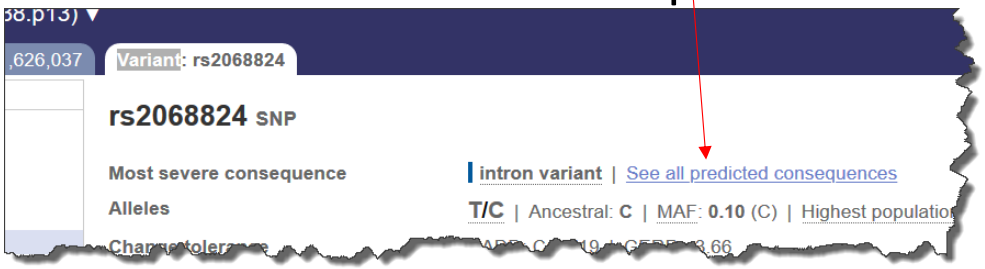
Finding The Phenotype Of A SNP Scenario *(cont-4)*

- What is Celiac Disease:
  - It is a serious autoimmune disease that occurs in genetically predisposed people where the ingestion of gluten leads to damage in the small intestine.



Finding The Phenotype Of A SNP Scenario *(cont-5)*

- On the “Variant rs2068824” tab click on the “See all predicted consequences” link.



  - This is an intron-variant
  - Scroll down to “**Gene expression correlations”** section. What tissues does this variant also show up in?
  - The effect size = the change in frequency of the SNP occurrence relative to the SNP on that page.
  - Click on the sample genotypes link on the upper left.
    - What is the number of subjects used to collect statistics about the genotype associated with this SNP?
  - On the “About this variant” line click the “This variant overlaps 2 transcripts“ link.

In-Lecture Exercise

- Suppose you are working with the sequence variation rs121908755

- What diseases is this variation associated with?
- What is/are the SNP(s) involved?
- Location of the variation?
- What are the GMAF data?
- What gene is involved?
- What other interesting info can you find out about this variation?

Next Generation Sequencing
Next Class

Extra Functional Genomics Material

Functional Genomics Technology Goals

- Generate sets of **full-length cDNA clones** and sequences that **represent human genes** and model organisms
- Support research on **methods** for studying functions of **non**protein-coding sequences
- Develop technology for comprehensive analysis of **gene expression**
- Improve methods for **genome-wide mutagenesis**
- Develop technology for **large-scale protein analyses**

http://www.ornl.gov/sci/techresources/Human_Genome/research/function.shtml

Definition (1) – Hieter & Boguski 1997

- The development & application of **global**
  - **Genome-wide or**
  - **System-wide experimental** approaches to assess **gene function** by making use of the **information** & **reagents** provided by **structural genomics**

- It is characterized by **high-throughput** or **large-scale** experimental methodologies
  - Combined with **statistical** or **computational analysis** of the results

Definition (2) – UC Davis Genome Center

- A means of assessing **phenotype differs** from more **classical approaches** primarily with respect to
  - **The scale & automation** of biological **investigations**
    - **A classical investigation** of gene **expression might examine how the expression of a single gene** varies with the development of an organism *in vivo*

- **Modern** functional genomics approaches, however, would examine **1,000-10,000 genes** are expressed **as a function of development**

http://genomics.ucdavis.edu/index_html.html

Definition (3) – Hunt & Livesey (ed.)

- x Subtracted cDNA libraries
- x Differential display (DD)
- x Representational difference analysis x
- Suppression subtractive hybridization x
- cDNA microarrays
- x 2-D gel electrophoresis

http://www.oup.co.uk/isbn/0-19-963774-1