Creating a Gene dossier

Your project is to compile a dossier about a gene that you choose and its corresponding protein. **Each student will create a Google website** on which to compile all materials related to the project. You will be graded on the materials compiled on your website and on an end-of-semester presentation of the project. The in-class final presentations of the project will use the website.

As we become familiar with bioinformatics tools, you will apply each of them to researching this gene/protein. Often, the information available from model organism studies is better than the information for the human version of the gene. When that is the case, fill in your dossier with information about the ortholog from a model organism, indicating in your notes which species the information came from. Use the Gene dossier assignment notes provided to you to record your text information as you go along. This will keep the format of your project notes consistent. You will also be saving other types of media, such as excel tables and images.

When you are asked to do cross-species comparisons, such as ortholog identification, limit yourself to information from the following list of 6 model organisms.

The 6 well-studied model organisms, with species names:

roundworm	*Caenorhabditis elegans*
fruit fly	*Drosophila melanogaster*
bacterium, Gram negative	*Escherichia coli*
human	*Homo sapiens*
budding (or bakers' and brewers') yeast	*Saccharomyces cerevisiae*
fission yeast (also a beer yeast)	*Schizosaccharomyces pombe*

Choose a gene from the list below. Although the names refer to human genes, there are orthologous genes in many other species, including model organisms. These are well-researched genes for which there is a wealth of information available.

Genes to choose from (2-3 students per gene):

gene	brief description
PSMC3	proteasome ATPase ring subunit
CDC20	ubiquitin ligase substrate adapter
DYNC1H1	dynein subunit, microtubule-dependent ATPase motor
CDK1	cyclin-dependent kinase
RAD51	recombinase for homologous recombination
MYH1	ATPase motor protein subunit
ACTB	β-actin
TCP1	CCT chaperone subunit
POLG	mitochondrial DNA polymerase
CDC5L	kinase required for the mitosis-to-G1 cell cycle transition
MTOR	signaling kinase, regulator of cell growth
RPPH1	RNase P RNA subunit
TUBB	β-tubulin
SPO11	DNA endonuclease, required for meiotic DSBs
POLD1	DNA polymerase subunit
CYCS	cytochrome C, mitochondrial electron transport
PPP2R4	protein phosphatase subunit
BLM	DNA helicase
MVK	mevalonate kinase, isoprenoid and sterol synthesis
CENPA	centromere-specific histone H3 variant
DHX9	RNA helicase
RPL28	ribosome component
HSP90	heat shock protein
EIF2S1	eukaryotic translation initiation
SIRT2	histone deacetylase
RAN	nuclear transport adapter
TERT	telomerase subunit
MAPK1	MAP kinase
CDC14A	Phosphatase, regulator of the cell division cycle
REC8	chromosome cohesion protein subunit
TOP2A	DNA topoisomerase II alpha

Part 1 General gene/protein description

Assignment: Gather basic information about the gene and its corresponding protein. Answer the questions as concisely as possible. Keep the headings that are in bold as a way to organize your information.

What is the gene's chromosomal location?

Is the gene spliced, and if so, how many splicing variants are known?

What is the function of the protein?

1 What is the individual protein's (or RNA's) biochemical activity?
Examples: cytoskeletal component, motor protein, transcription factor, histone, polymerase, helicase, kinase, phosphatase.

2 What is the protein's (or RNA's) cellular or biological role?
Examples: vesicle transport, cell cycle regulation, activation of metabolic genes, protein synthesis, chromosome segregation.

What is the subcellular localization of the protein?

Is the protein a member of a stable complex?
State whether the protein functions individually or as a member of a complex. If it functions in a complex, briefly describe the complex.

Find a structural image of your protein if one is available.
The structure might be a crystal structure, cryo-EM, NMR…etc.

What post-translational modifications of the protein are known?
Be as specific as possible about what amino acid residues are modified and what type of modifications they are.

Resources:
NCBI Gene, model organism websites, such as Saccharomyces Genome Database (SGD), Flybase, Pombase, Wormbase, EcoCyc, current literature.
Other useful websites: UniProt, MIPS

Part 2 Paralogs and orthologs

Assignment: Identify the paralogs of your protein and its orthologs (if they exist) In the 6 model species we are using. Use BLAST to align the protein sequences of the orthologs. For your alignment, choose one ortholog (if available) from each the six model species and highlight conserved residues or regions of the protein.

Tables of paralogs and orthologs:
Make 2 excel tables, one titled "paralogs" and one titled "orthologs". Consult the definitions of paralog and ortholog if necessary.

Paralogs table
The paralogs table will be a simple list of the paralogs in a single column with your gene at the top.

Orthologs table
The orthologs table will have several columns, organized as follows:

Column 1 titled "species", will list the 6 model species.
Column 2 titled "ortholog 1", will list the closest ortholog in each species
Column 3 titled "ortholog 2", will list an additional ortholog from each species, if one is known.

Don't go beyond 2 columns of orthologs.

Protein sequence alignments
Get the amino acid sequences of the nearest orthologs of your protein from each model species. Align the sequences using BLAST and save a figure of the alignment to use in your end-of-semester presentation.

Resources:
P-POD ortholog identification tool, InParanoid ortholog identification tool, UniProt, NCBI BLAST

Part 3 Functional profiles (phenotypes) and disease-causing mutations

Assignment: Make a functional profile of your protein by listing the phenotypes of mutations in the gene. Examples of interesting phenotypes include DNA damage-sensitivity, inability to utilize metabolites, defects in cell morphology, human diseases, etc. If human disease alleles are not known, identify mutations in model organisms that might be informative about human diseases.

More detailed instructions:
Make an excel table titled "phenotypes and disease alleles". List the most interesting phenotypes first. Stop at a maximum of 12 mutant alleles and 3 phenotypes per allele. Highlight in yellow the rows that describe human disease phenotypes.

Have the following columns:

Col 1 "species"	the species in which the mutation was identified
Col 2 "gene"	the gene name, which may differ depending on species
Col 3 "mutation"	the nucleotide change(s) in the gene, if known
Col 4 "aa change"	the amino acid change(s) in the protein, if known
Col 5 "phenotype 1"	a brief description of the phenotype
Col 6 "phenotype 2"	a brief description of a second phenotype
Col 7 "phenotype 3"	a brief description of a third phenotype

If possible, include up to 6 human disease alleles of your gene. Choose the "best" or most pathogenic alleles if possible. If phenotypes for your human gene are few or poorly-characterized, list the phenotypes of mutations in orthologous genes, ideally in conserved amino acid positions.

Some hints for OMIM:
Search for your gene (or its human ortholog) in NCBI OMIM, and scroll down the page to find the "Variation" heading. Click "See variants in ClinVar". You will find a list of mutations in your gene, indicating the specific DNA sequence change(s) and the corresponding change (or no change) in the protein sequence.

Resources:
NCBI Online Mendelian Inheritance in Man (OMIM), primary literature, model organism websites

Part 4 Networks

Assignment: Use STRING to map the known interactions of your gene/protein with other genes/proteins. If the protein is in a complex, list the other members of the complex. For *E. coli* genes, operon structure can be a big clue to a functional relationship.

Produce the following graphs:
Network of physical (protein-protein) interactions
Network of genetic interactions
Can the genetic interactions be graphed as positive vs. negative?
For *E. coli* genes, is the gene in an operon? If so, prepare a figure of the operon as it is organized on the bacterial chromosome.

Reading:
Szklaryczyk et al., 2015. STRING v10, protein-protein interaction networks, integrated over the tree of life.

Resources:
The STRING database
BioGrid

Part 5 Experiment design

Assignment: Design an experiment using systems biology tools we have studied. Use the following subheadings to organize your experiment:

Hypothesis
(address an outstanding question or gap in our knowledge)

Methods and experimental design

Data analysis

Examples:
Could you design an experiment to measure the interactions of your protein with other proteins or DNA under some specific condition that could be very informative, such as during embryonic development, during heat shock stress, in cancer vs. normal cells, etc.

Two heads are better than one when brainstorming experiments, so work in groups of 3 students that are studying the same gene.

Part 6 Presentation

Assignment: Turn in your gene dossier as a single web page and email the URL to the TA AND PROFESSOR.

Presentations:
Each student group (2-3 students working on the same gene) should prepare 5-8 figures as Powerpoint slides to illustrate important parts of the project. We will project the figures during your presentation, which will last 7-10 minutes, depending on the number of groups. Some suggested figures: structural images of your protein, alignments of your protein to orthologous proteins, networks of physical and genetic interactions, figures of experimental designs, etc. I especially encourage presenting on the experiments you suggest as a group.