Statistics in Health Science

Laura Saba, PhD
Assistant Professor
Department of Pharmaceutical Sciences

Skaggs School of Pharmacy
and Pharmaceutical Sciences

100+ years of education, patient care & scientific discovery.

Overview

- What is Biostatistics?
- 3 quick tips to become statistically literate
- Examples of advanced biostatistics

WHAT IS BIOSTATISTICS?

Definition of Statistics

- **Statistics** is the science of collecting, analyzing, and drawing conclusions from data in a sensible way.

  Statistics offers us powerful tools for gaining insight into the world around us.

  "It is the mark of a truly intelligent person to be moved by statistics" - George Bernard Shaw

MONEYBA
BASED ON A TRUE S
ORIGINAL MOTION PICTURE SOUN
MUSIC BY MYCHAEL D
INSPIRADA EN UNA HISTORI
21
BLACKJACK
the signal
and the noise
why so many
predictions fail—
but some don't
nate silver
NUMB3RS

> “The best thing about being a statistician is that you get to play in everyone else’s backyard.”

John Tukey, Bell Labs, Princeton University

> “I love that statistics is very multi-disciplinary. It involves problem solving in a group environment and it involves many skills and talents. I love the ability to be a mathematician, computer scientist, teacher, quizmaster, sleuth, and devil’s advocate all rolled into one.”

Linda Quinn, Private Industrial Consultant

> “Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”

Samuel S. Wilks

Wikipedia Definition of Biostatistics

- **Biostatistics** (or **biometry**) is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine, pharmacy, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results. A major branch of this is medical biostatistics, which is exclusively concerned with medicine and health.

Fields of Application

- Biomedical research
- Animal health
- Pharmacology
- Genetics
- Chemistry
- Epidemiology

Education

- **High School**
  - Study statistics, mathematics, science, computer science, and English
- **College**
  - Few undergraduate programs in biostatistics; many majors are relevant if thinking about biostatistics as career (e.g., statistics, biology, chemistry, computer science)
- **Post-Graduate**
  - Many biostatistics jobs require a Master’s degree or PhD

Skills

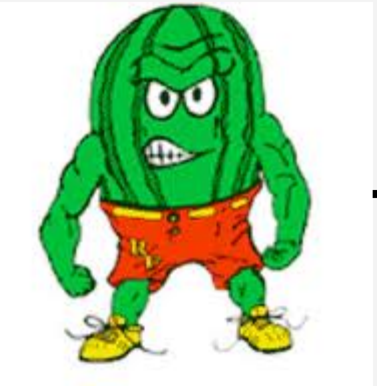
- **Quantitative Skills**
  - Statistics, Mathematics, Science
- **Problem Solving Skills**
  - Analysis, Teamwork
- **Communication Skills**
  - Verbal, Written
- **Computer Programming Languages**
- **Foundation in Field of Application**

Why a career in (bio)statistics?

Taken from http://www.biostat.ucla.edu/why-biostat

1. The heavy lifting is intellectual, not physical.
2. The work is indoors.
3. It pays well. Starting salaries for a PhD in Biostatistics often exceeds $100,000 per year (“For Today’s Graduate, Just One Word: Statistics”, Steve Lohr, NYT, August 6, 2009)
4. The skills are transferable and attractive to employers.
5. The work is collaborative.
6. It is rewarding to solve real life problems, using skills that few people have. Our statistical models can help cure diseases and improve quality of life.
7. It is intellectually stimulating!

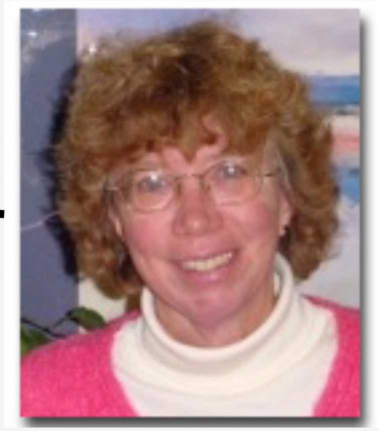
My Journey



Mighty Meloneer
Rocky Ford, CO

Oredigger
Chemistry Degree
from CSM

??
PhD in Biostatistics



Diane Fairclough
“Latent Pattern
Mixture Models”

10 years studying
genetics for
alcohol-related
traits

3 QUICK TIPS FOR STATISTICAL LITERACY

Number of children per family example

Inference

- An **inference** is a conclusion that patterns in the data are present in a larger context.
- **Statistical inference** is an inference justified using statistical methods.
- **Causal inference** is drawing a cause and effect relationship between an explanatory variable and a response variable.
- **Scope of inference** is the group of objects to which my conclusion (results) extend.

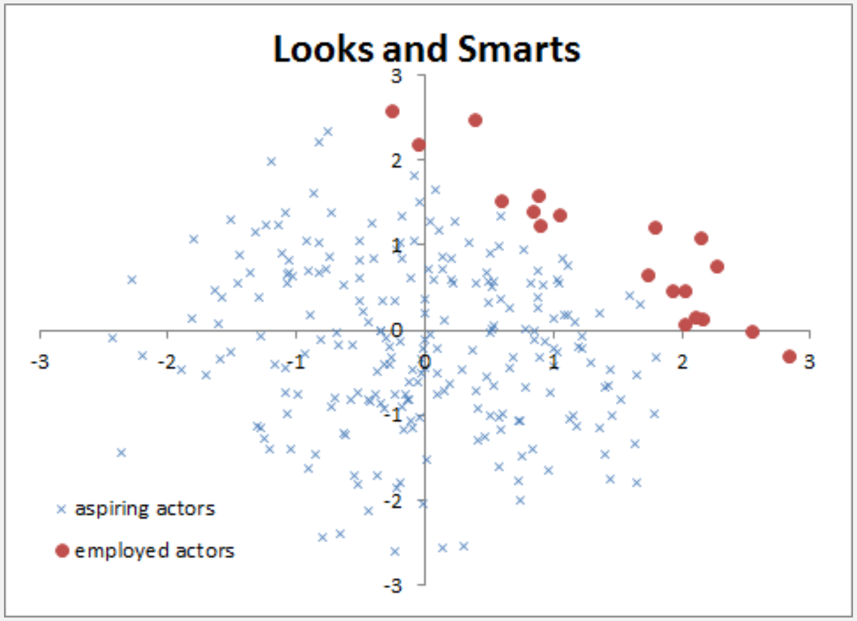
SCOPE OF INFERENCE

Populations and Samples

Researchers are typically interested in finding results that apply to an entire population of people or things.

- Population
  - » The collection of units (be they people, plankton, plants, cities, suicidal authors, etc.) to which we want to generalize a set of findings or a statistical model.
- Sample
  - » A smaller (but hopefully representative) collection of units from a population used to determine truths about that population.

- If good looks and smarts are distributed normally,
- If good looks and smarts **have nothing to do with each other**
- If movie producers want both smarts and looks Then, by observing **employed actors** we'll assume that looks and smarts have a **negative correlation**



http://euri.ca/2012/youre-probably-polluting-your-statistics-more-than-you-think/index.html
http://www.theatlantic.com/business/archive/2012/05/when-correlation-is-not-causation-but-something-much-more-screwy/256918/

Scope of Inference

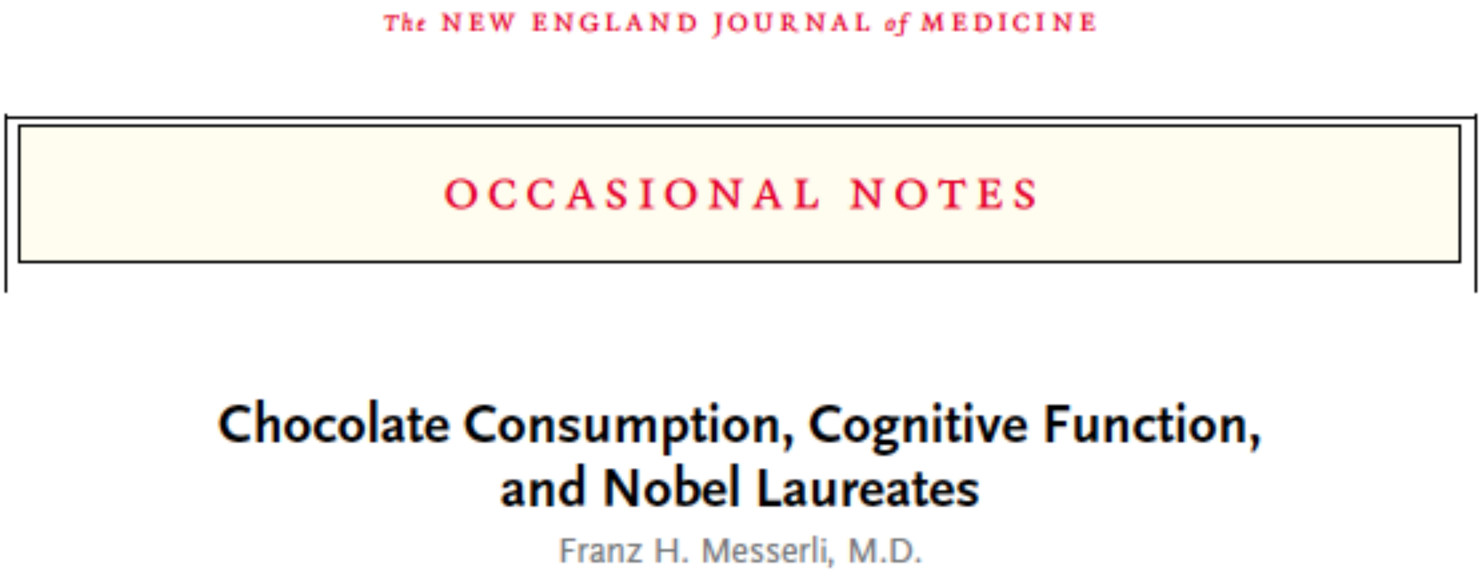
The **scope of inference** is the group of individuals to whom the statistical conclusions can be extended.

- Inferences to populations can be only drawn from random sampling studies.
  - A **random sampling** study is when units are randomly selected from a well-defined population.
  - Random sampling typically ensures that all subpopulations are represented in the sample in roughly the same proportion as the population.
  - Our statistical procedures take into account that sometimes the sample may not be a very good mix of the population.

Scope of Inference (cont.)

- When subjects are not obtained through random sampling the results or model extend to the sampled group but not the larger population.
- The most basic form of random sampling is **simple random sample**.
  - A simple random sample of size n from a population is a subset of the population consisting of n members selected in such a way that every subset of size n has the same chance of being selected.
  - List every unit in the population and randomly select n of the units.

CAUSAL INFERENCE



The NEW ENGLAND JOURNAL of MEDICINE

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

"Switzerland was the top performer in terms of both the number of Nobel laureates and chocolate consumption. The slope of the regression line allows us to estimate that **it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1**. For the United States, that would amount to 125 million kg per year. The minimally effective chocolate dose seems to hover around 2 kg per year, and the dose–response curve reveals no apparent ceiling on the number of Nobel laureates at the highest chocolate-dose level of 11 kg per year."

Types of Studies

- **Experiments** are studies in which we manipulate one or more explanatory variables (e.g., treatments) to see the effect they have on another variable.
  - In a **randomized experiment** the investigator uses a chance mechanism to assign experimental units to various treatment groups
- **Observational studies** are studies in which the data are measured through observation of the world as it naturally occurs.
  - Grouping (i.e., explanatory variable) occurs naturally and is not assigned

Causal Inference to Observational Studies

- Causal inference is impossible in observational studies because confounding variables may cause the differences in the behavior of the response variable for different groups.
  - » A **confounding variable** is a variable that explains the group a person is in and also the outcome of interest.
    - e.g., health consciousness is a confounding variable when testing the relationship between takes vitamins and how often a person gets sick
    - e.g., smoking is a confounding variable when testing the association between coffee consumption and lung cancer

Causal Inference in Randomized Studies

- Causal inference can be made from randomized experiments but not from observational studies.
  - Randomization ensures that subjects with different features (i.e., confounding variables) are mixed up evenly among the treatment groups.
  - Randomized experiments seek to create groups that are totally similar except whether a treatment is present or absent.
  - The possibility that the groups may not end up being very “random” (i.e., groups are not mixed very well) is incorporated into the statistical tools used to express our uncertainty.

INTERPRETING P-VALUES

Null and Alternative Hypothesis

- Null Hypothesis
  - Half of the deck of cards are red
- Alternative Hypothesis
  - The is an unequal number of red and black cards in the deck

P-value

- P-value: the probability that chance alone leads to a test statistic as supportive or more supportive of the alternative hypothesis as what we saw in our sample if the null hypothesis is true.
  - The probability of seeing ten or more red cards if half of the cards in the deck are red.
  - IT IS NOT: the probability of the deck not having an equal number of red and black cards

QUICK REVIEW

When can we generalize results from a sample to a population?

- Results from a sample can only be generalized to a population when units are **randomly selected** from a well-defined population

Under what conditions can we make cause-and-effect statistical conclusions?

- Only in a **RANDOMIZED EXPERIMENT**
- For Bonus Points - Why?
  - When an independent variable (often treatment group) is randomly assigned to each subject, you assure that the value of the independent variable is not related to any other characteristic of the subject. (i.e., you avoid confounding)

Definition of a P-Value

- the probability that chance alone leads to a test statistics as supportive or more supportive of the alternative hypothesis as what we saw in our sample if the null hypothesis is true

1. probability that chance alone leads to a test statistics as supportive or more supportive of the **alternative hypothesis**
2. as what we saw in **our sample**
3. if the **null hypothesis** is true

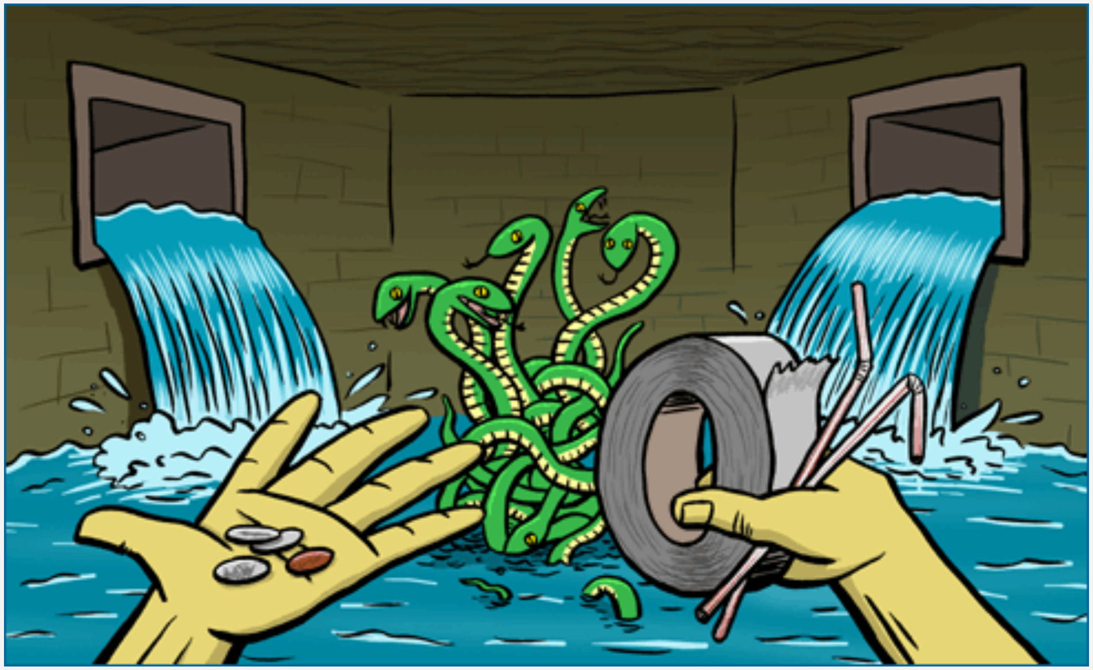
EXAMPLE OF ADVANCED STATISTICS

Systems Genetics

“...the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models" (Sauer et al., *Science 2007*).

“Good Enough Solutions”

Weiss et al (2012). “Good Enough Solutions” and the Genetics of Complex Disease. Circ Res 111:493-504.



http://evolution.berkeley.edu/evolibrary/article/mantisshrimp_01

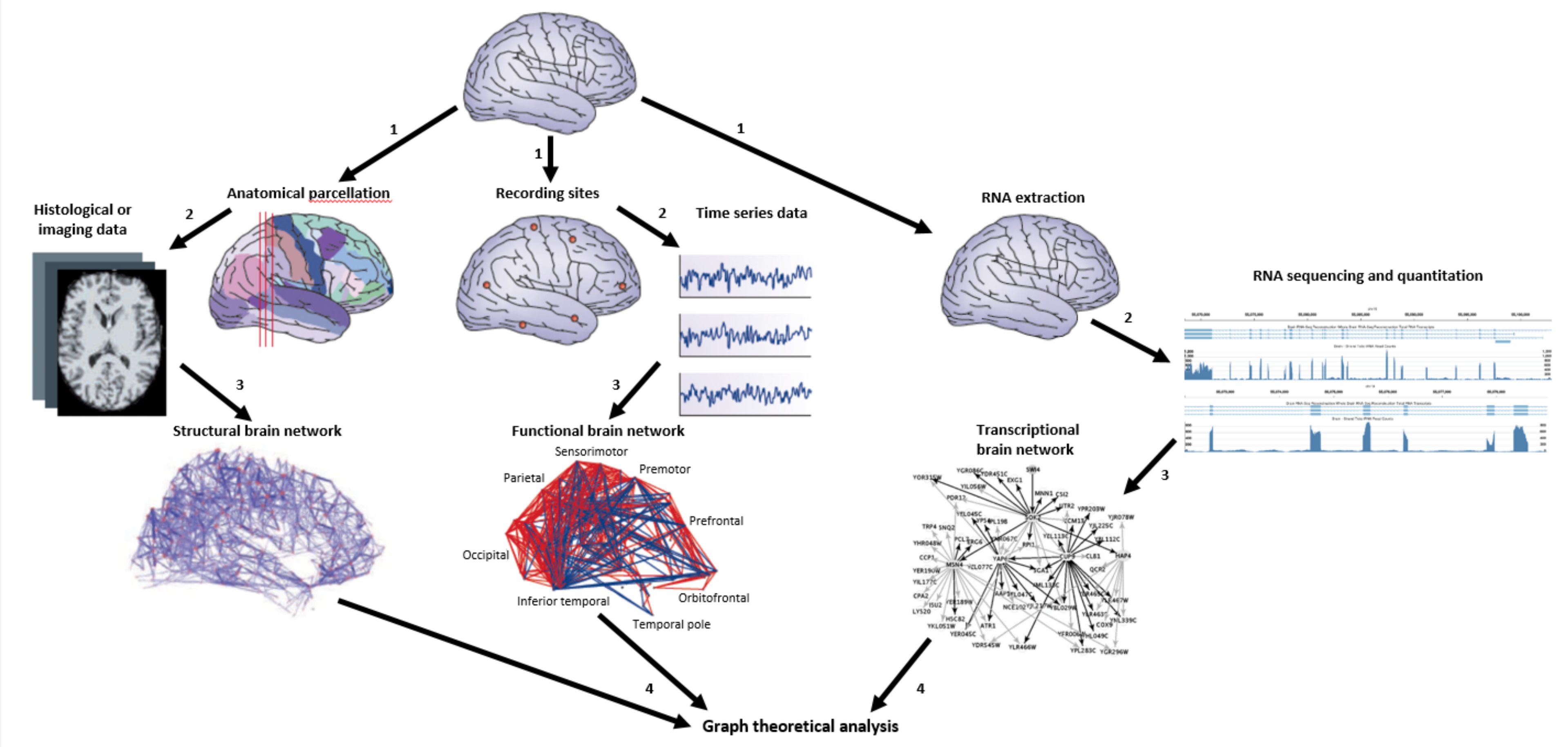
- Concept from evolutionary biology
- “in complex systems, many different combinations of the system’s parameters can produce a nearly identical output”

Good Enough Genetic/Genomic Solutions

- DNA sequence and RNA expression levels = ‘system parameters’
- Genetic diversity → differences in baseline systems
- Differences in baseline systems → differences in response to environmental variables

Can we statistically describe these baseline systems?

Brain Transcriptional Connectome



Why Study the RNA Dimension

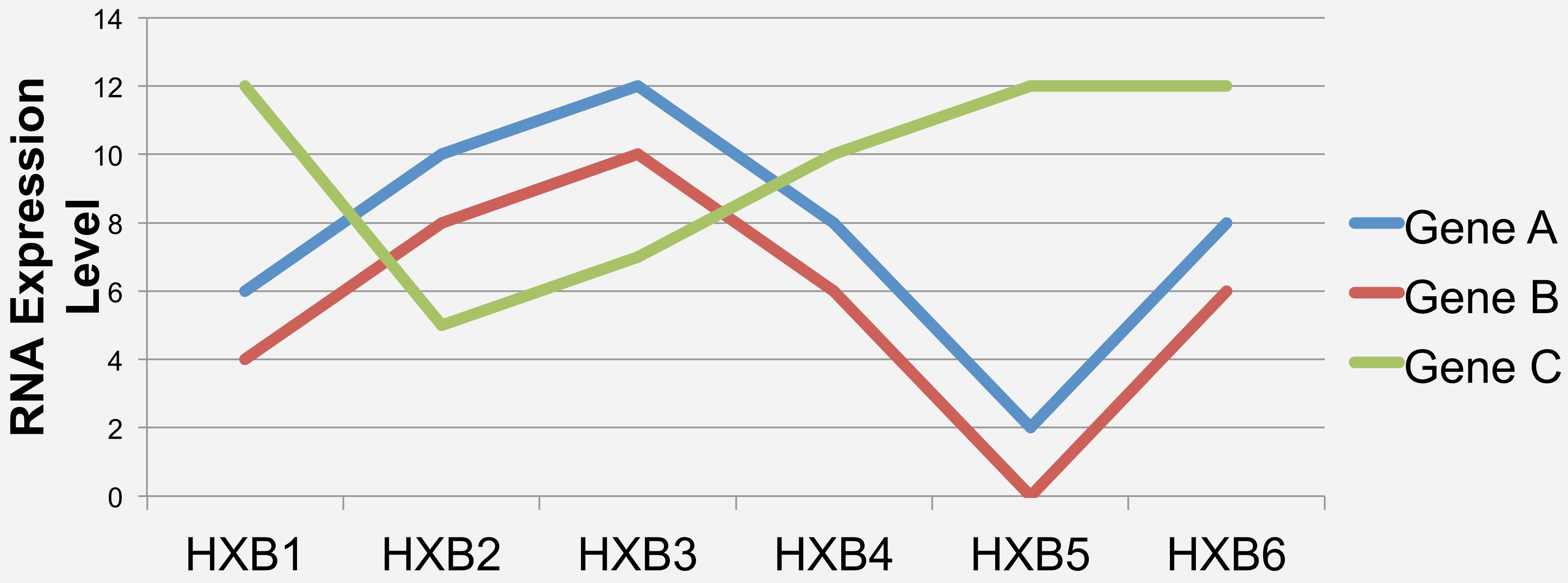
Transcriptome links DNA and complex traits/diseases

A. One of the first quantitative links between DNA sequence and phenotype

B. First step where DNA sequence and environment interact

C. Implementation of graph theory at the transcript level provides insight into genetic/environmental interactions that are the basis for susceptibility to complex diseases.

Co-expression as a measure of the "connectome"



Theory – if the magnitude of RNA expression of two transcripts correlates over multiple "environments" (genomes), then the two transcripts are involved in similar biological processes

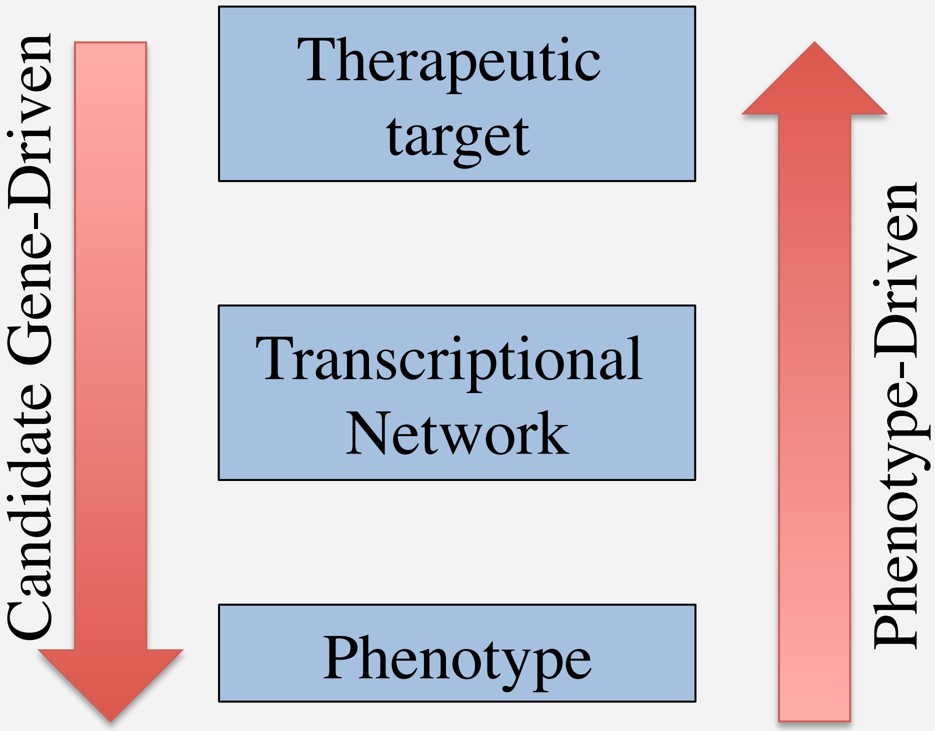
Weighted Gene Co-Expression Network Analysis

Why Not Just Use Correlation?

1. Simple correlation does not give connectivity.
2. How are we measuring co-expression?
   - » Scale-Free Network
     - Network has few highly connected genes rather then each gene have similar connectivity
     - **Biologically motivated**, fewer highly connected genes means that a system is more robust to failure of any one gene
3. How do we get a **robust** measure of connectivity for identifying modules?
   - » Topological Overlap Measure
     - Includes a measure of how many "friends" two genes have in common
     - Protects against spurious correlations among genes

Two Primary Approaches To Using the Transcriptional Connectome

1. **Candidate gene-driven** analysis of biological/ genetic context
2. **Phenotype-driven** genome-wide analysis for candidate genetic pathways for predisposition to disease

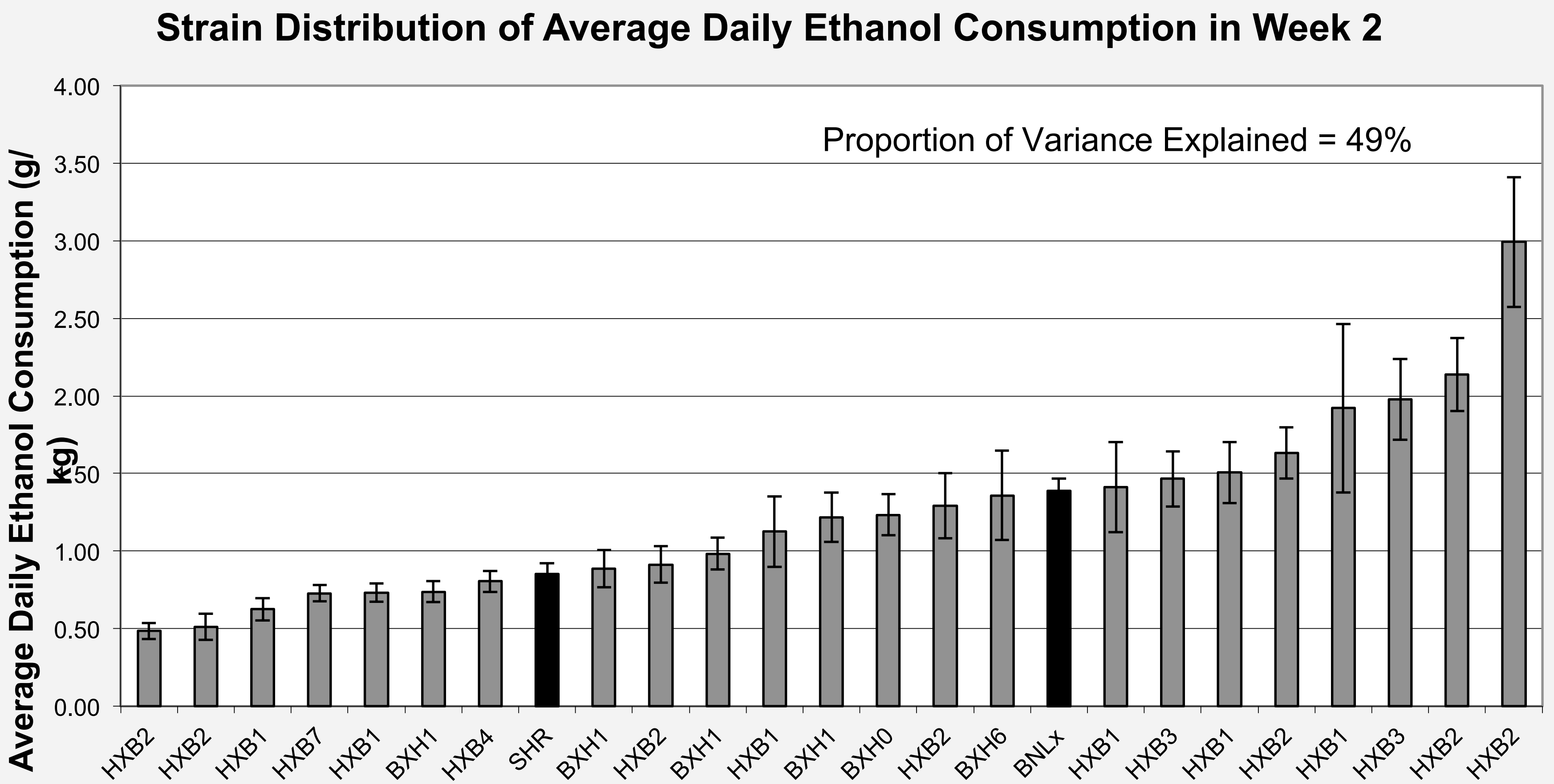


Genetics and Alcohol Consumption

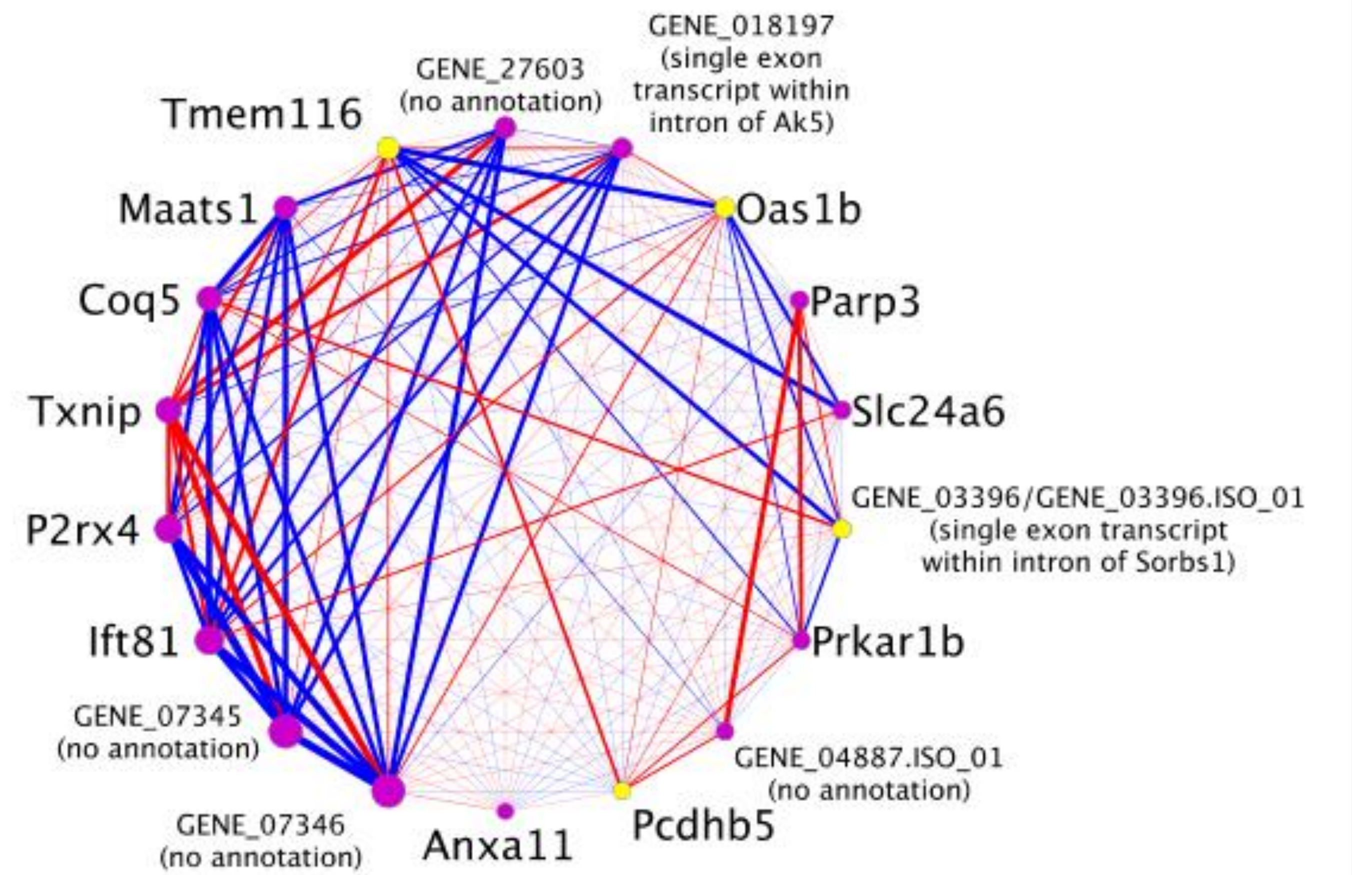
- etiologic essential
- strong genetic influence
- complex polygenic trait

http://lets-go-to-the-movies.tumblr.com/tagged/Bridesmaids/page/4

Alcohol Consumption Across a Genetically Controlled Rat Population

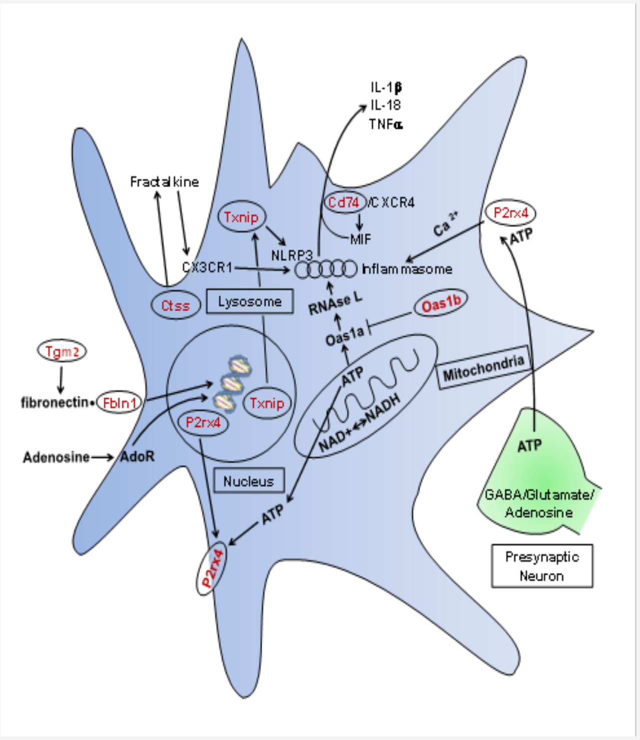


Coexpression Module Associated with Alcohol Consumption

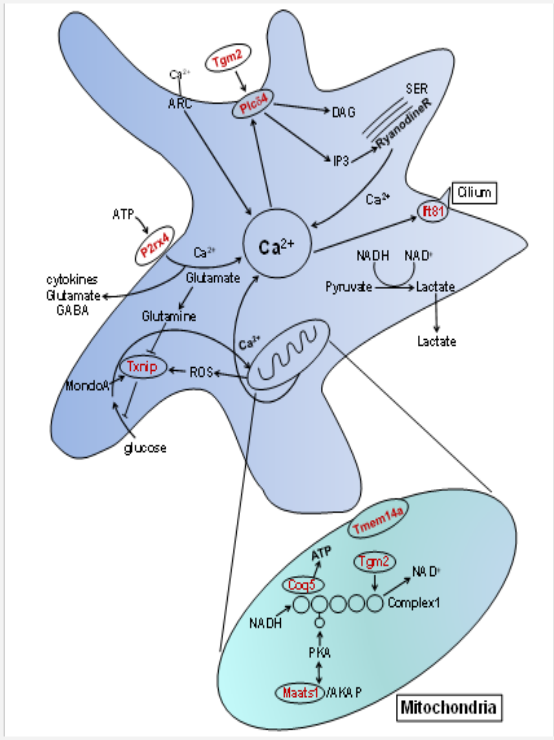


Saba et al (2015). The sequenced rat brain transcriptome, its use in identifying networks predisposing alcohol consumption. FEBS (in press).

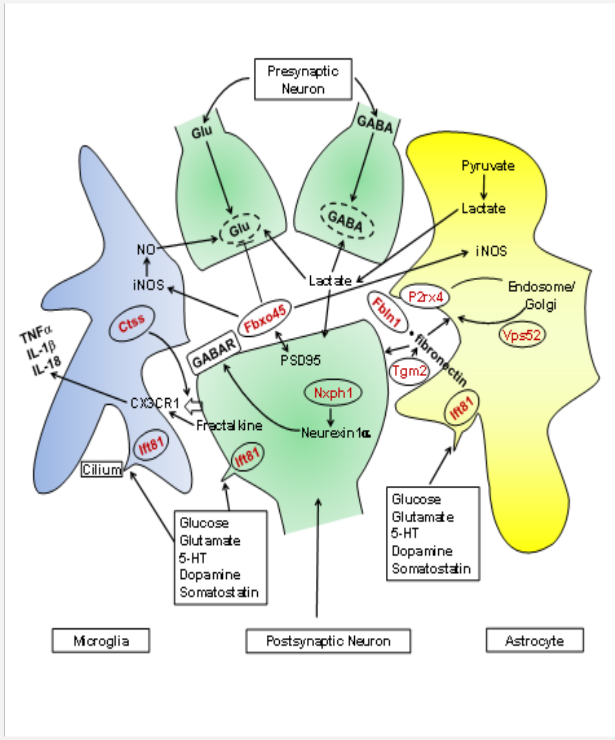
Biological Context from Pathway



Inflammation/Immune Response



Energy/Ca2+ Homeostasis/ Redox

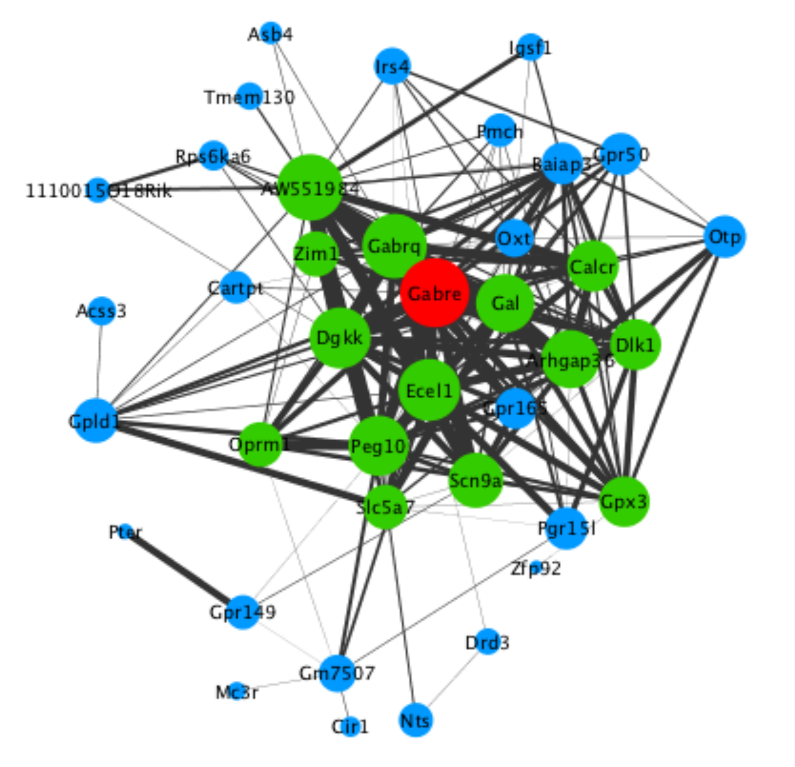


Glial/Neuronal Communication

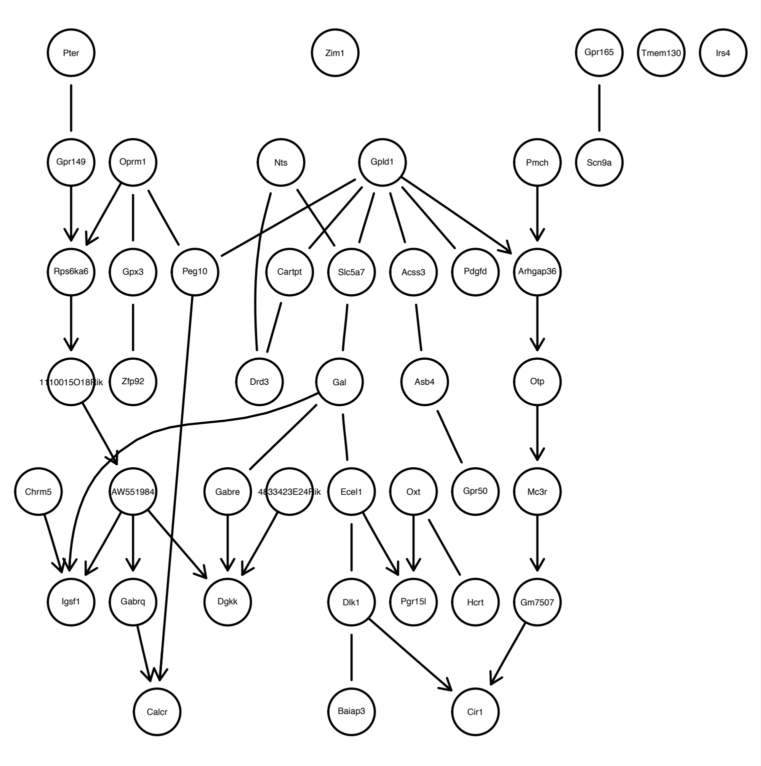
Saba et al (2015). The sequenced rat brain transcriptome, its use in identifying networks predisposing alcohol consumption. FEBS (in press).

FUTURE RESEARCH

WGCNA

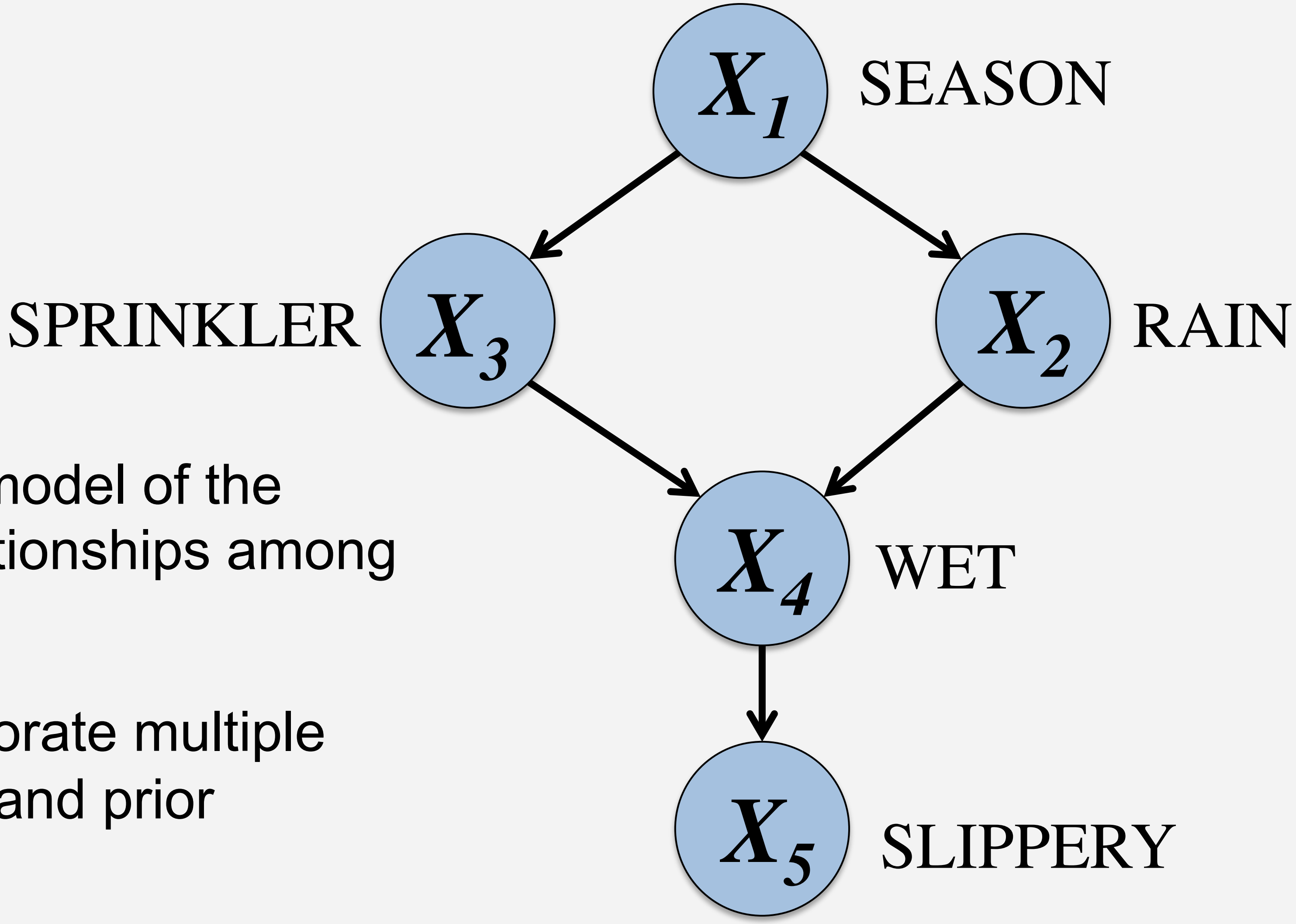


Bayesian Networks



Moving from Description to Prediction

Bayesian Network Models

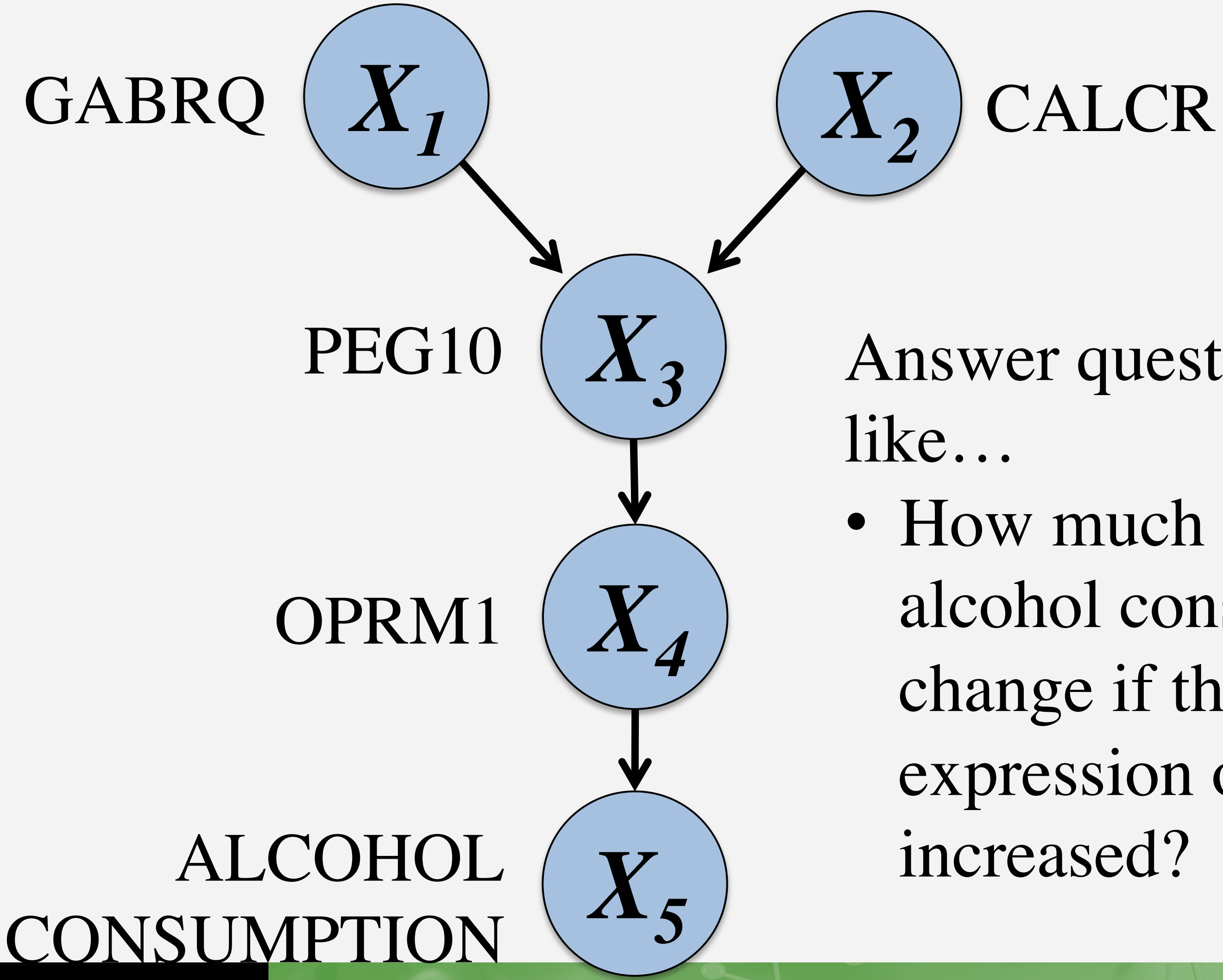


- Graphical model of the causal relationships among variables
- Can incorporate multiple data types and prior knowledge

Reasoning with Bayesian Networks

- With new information (e.g., new observation or proposed pharmacological manipulation) these models can be used for **probability propagation**.
  - Diagnostic reasoning – what is the probability that it rained if the ground is slippery?
  - Predictive reasoning – what is the probability of the sidewalk being slippery if it rained last night?

Bayesian Networks in Drug Target Identification



Answer questions like…

- How much will the alcohol consumption change if the expression of Calcr is increased?