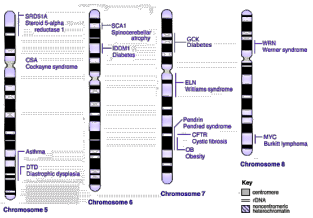
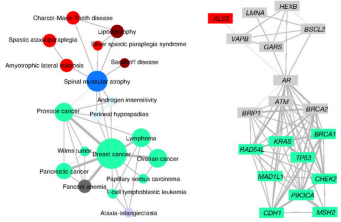
# Genes and Diseases

DS2001, Prof. Rachlin







## Background

In this practicum, we will explore GAD, the Genetic Association Database (Becker *et al.*, 2004). In the process we'll discover links between two seemingly unrelated conditions or diseases. Some diseases are associated with a single mutation of a single gene. But most diseases or conditions are "multi-genic" -- many genes have been positively linked with the disease, though the nature of the connection may not be well-understood.

While cataloging these associations is an important first step in drug discovery, it is important to remember:

- Identifying that there is a connection between a gene and some disease does not mean that the underlying biological mechanisms are well-understood.
- The association may have been tied to a particular sub-population (Japanese, American Indian, etc.)
- The association may only suggest some increased *probability* of acquiring the disease. There may be other, yet unknown connections with environmental factors such as diet which have yet to be teased out.

## The GAD Data

The GAD data set I have given you (gad.csv) has the following columns:
**gene** – The official gene symbol
**disease** – The name of the disease or "phenotype"
**pubmed_id** – Publication identifier (PubMed)
**year** – Year of publication

## Instructions

For your convenience, starting code has been provided.

1. Find all genes linked to Asthma
2. Figure out which disease (other than asthma) is also linked to the greatest number of asthma-linked genes.
3. Google Asthma and the name of your disease. Does research suggest a biological connection between these two diseases? In your submission, provide an example URL.
4. **To earn a 5**: Survey all diseases. For each disease find the disease having the most gene overlaps. You may want to limit your result to pairs of disease that have at least 10 or 15 overlapping gene associations.

## Submit

Code (gad.py) and program output(s) as text files.