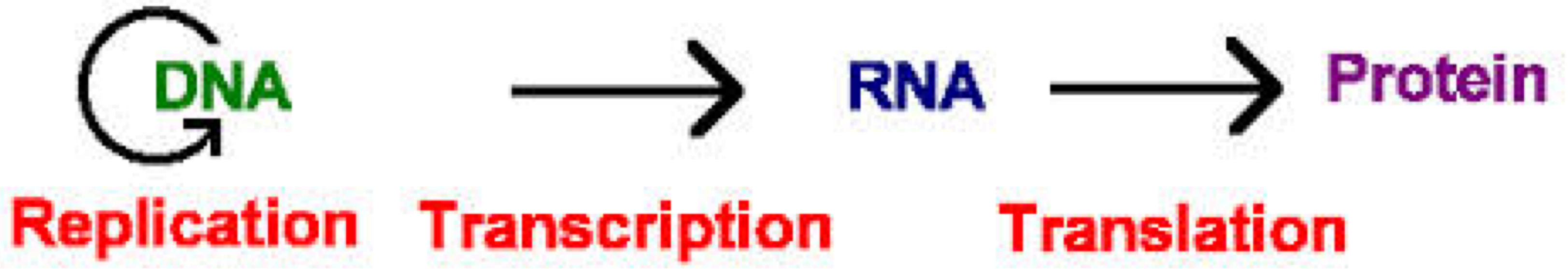
Integrative Analysis

Alison Motsinger-Reif, PhD
Branch Chief, Senior Investigator
Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences

alison.motsinger-reif@niehs.nih.gov

Integrative Analysis

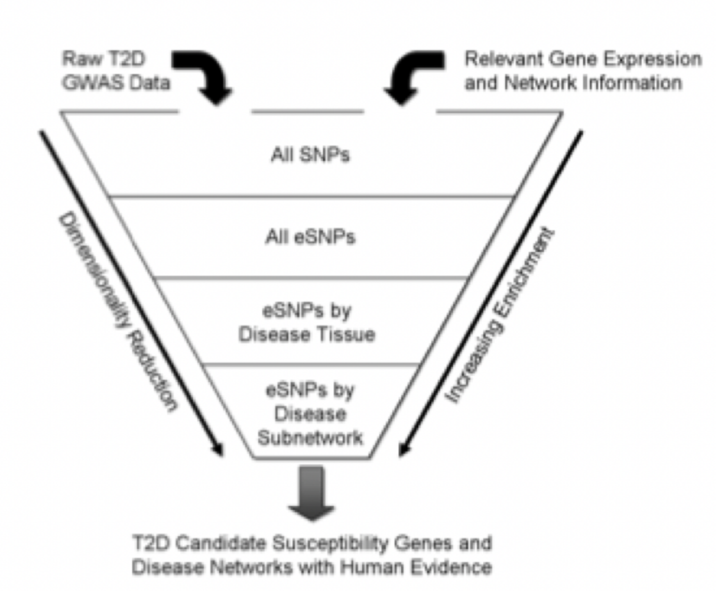
- Many motivating reasons to combine/integrate data from multiple “-omes”
- Expression and SNP data is most commonly done
  - Though methods could be applied to combine other “-omics”
- Generally make assumptions about central dogma



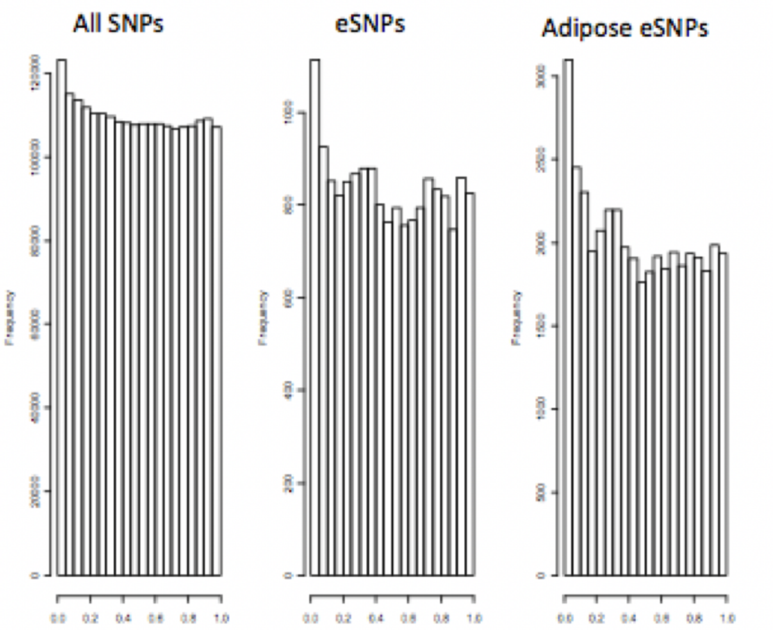
…… though we are learning more and more examples of exceptions to this….

Filtering

- Trade-off between unbiased discovery and improving power
- Expression and SNP data is most commonly done
  - Though methods could be applied to combine other “-omics”
- Pathway analysis in one –ome can narrow hypothesis tests in other(s)
- Still need to correct for multiple testing


Raw T2D GWAS Data
Relevant Gene Expression and Network Information
All SNPs
All eSNPs
eSNPs by Disease Tissue
eSNPs by Disease Subnetwork
Dimensionality Reduction
Increasing Enrichment
T2D Candidate Susceptibility Genes and Disease Networks with Human Evidence

Zhong et al. (2010) Elucidating Networks of eSNPs associated with Type 2 Diabetes.



Considerations: Multiple Test Correction

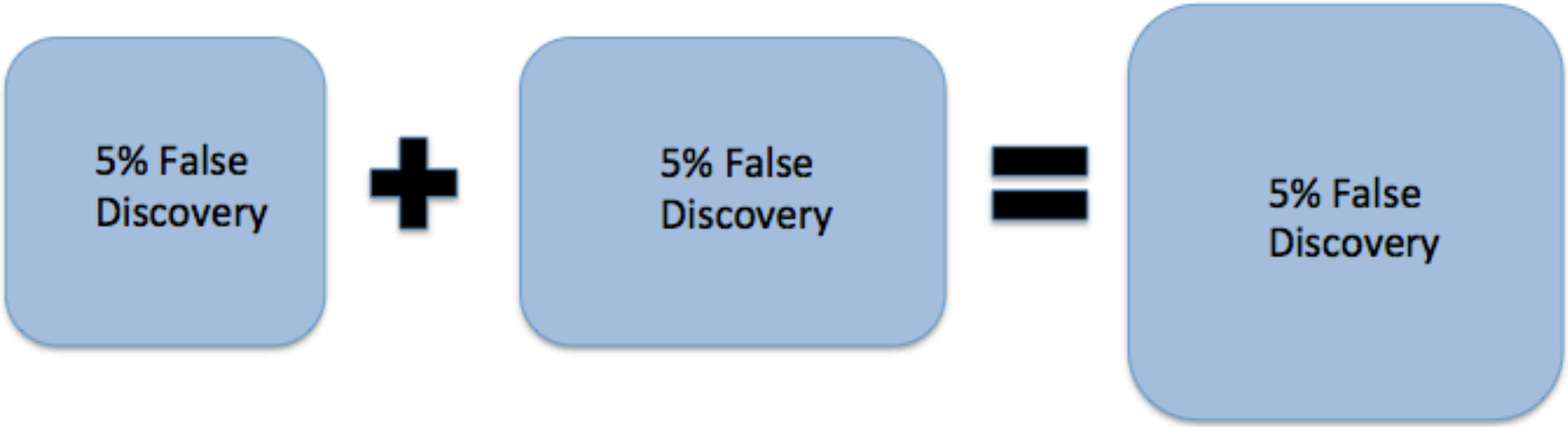
- Can be valid to test hypotheses in a partitioned fashion if:
  1. The partitions are specified **before** you look at the data
  2. Your multiple testing procedure controls the overall error rate

5% P-value vs 5% FDR

- P-value -> Over a large number of times the experiment is repeated, 5% of the time we'll identify 1 or more false positive SNPs
- FDR -> 5% of identified SNPs are false positives

Partitioned Testing (FDR)

- Simple way to control error over multiple partitions
- Controlling FDR at level ξ in each (non-overlapping) set, results in overall FDR ξ



5% False Discovery + 5% False Discovery = 5% False Discovery

eSNPs: Computing your own

- eSNP analyses are just GWAS's with continuous traits, but 1000's of them
- Approaches:
  - Frequentist:
    - Linear Regression
      - Outlier sensitive, can adjust for covariates
    - Robust Regression
      - Outlier resistant, can adjust for covariates, more computationally demanding
    - Kruskal-Wallis
      - Nonparametric (outlier resistant), difficult to adjust for covariates
  - Bayesian:
    - More resistant to outlier effects than linear regression, but require setting priors on each parameter
    - Some software available:
      - Bimbam
      - SNPTEST

eSNPs: A note on computation

- eSNP analysis is extremely resource intensive in both processor time and storage
- Computation requires a cluster (not possible on a desktop machine)
- Storage: $N_{markers}$x $N_{expression\ traits}$ is typically large
  - One approach is to store only results with pvalue < some threshold

eSNP Discovery

- eSNPs near gene location are easier to find
  - Real biological effects (*cis* regulation)
  - Fewer hypothesis tests relative to genomewide
- Typical approach is to identify local (proximal) eSNPs and distant (distal) eSNPs in separate steps
- Controlling each at fixed FDR, ξ, controls the overall FDR at ξ
- Choice of proximal window can effect eSNP discovery

eSNPs: Publicly available

- Databases:
  - www.scandb.rog
  - http://eqtl.uchicago.edu/gbrowse/eqtl/
  - https://gtexportal.org/home/

- Emerging number of tissue specific resources:
  - Harvard Brain
  - Kronos Phase 1- Brain, Alzheimers
  - Human Liver Cohort
  - ….

Integrated Analysis

- Newer approaches will allow you to not do partitioned/filtered analysis, and leverage information across datatypes

- New technologies allow for more ready integration
  - Ex. RNA-Seq
  - Dropping costs allow for more datatypes to be collected simultaneously
  - Biobanking effort are storing more tissues

Motivation for Integrated Analysis

- Naturally allow Bayesian approaches for identifying priors or jointing modeling data

- Several new approaches proposed
  - Methods that were developed for eSNPs are readily extended across data types
  - Other approaches take into account similarities between/withing phenotypes
    - Several an ontology jointly representing disease risk factors and causal mechanisms based on GWAS results
    - Proposed ontology is disease-specific (nicotine addiction and treatment) and only applicable to very specific research questions

Summary on Integrated Analysis

- Database development, curation, editing, etc. always lags behind technology
- Issues with incomplete and inaccurate annotation accumulate as more “omes” are considered
- With more complex data, this complexity is not readily captured in the databases the gene set analysis relies on
  - Differences in cell types, exposure, time, etc.
  - Major needs for methods development…..

Questions?