

# **Practical Considerations in Genomic Research**

**Yvette P. Conley, PhD  
Associate Professor  
School of Nursing  
University of Pittsburgh**

# So you want to do an association study...

- **Think about who to recruit/genotype**
  - Related or unrelated subjects
    - Often depends on disease and availability of parents and family members
  - Controls
  - Cohorts available
- **Give thought to phenotyping**
  - This includes “unaffecteds” too
  - Traditionally how is phenotyping conducted for the condition/trait of interest (can help with meta-analyses)
  - State-of-the-art, gold standard phenotyping
- **Candidate gene(s) vs Genome Wide**

# Candidate Gene vs Genome Wide Association Study?

- A candidate gene association study is used when you have an educated guess about the involvement of a gene(s) in the condition you're investigating
- A Genome Wide Association Study (GWAS) is used when you want to investigate the entire genome to determine what variant(s) are associated with the condition you're studying
  - One advantage is the underlying biology of the condition does not need to be well understood

# So you want to do a candidate gene association study...

- **Justify the gene(s) for investigation**
  - provide biological plausibility
    - etiology of the condition
    - treatment of condition
    - pathway related genes
  - expression of gene(s)
  - positional candidates

# So you want to do a candidate gene association study...

- **Select polymorphisms to evaluate in/around the gene(s)**
  - tagging SNPs
    - example: IL6 has 158 entries in dbSNP for humans
    - tSNPs for IL6 using  $MAF \geq 20\%$ ;  $r^2 \geq .80$ ; CEU is 2
  - functional polymorphisms
    - provide evidence to support it being functional
    - evidence does not need to be in your population – simply that its function has been documented
    - non-synonymous polymorphisms are potentially functional if can't provide specific evidence

# So you want to do a genome wide association (GWA) study...

- **Justify use of this approach**
- **Requires much larger sample sizes**
- **CNV**
- **Cost**
  - While per genotype cost is extremely low – per subject cost is high due to # of genotypes generated → however good news is the cost keeps coming down!
- **Publicly available GWAS data**

**Table 2.** The table shows the power that can be achieved by each chip with a total budget of \$2,000,000.

<b>Chip</b>	<b>Average Price (\$)</b>	<b>Number of cases/controls</b>	<b>Power</b>
Affy500 k	420	2381	0.767
Illu300 k	377	2653	0.821
Illu610 k	452	2212	0.818
Affy6.0	505	1980	0.772
Illu1M	750	1257	0.635
Complete	-	2653	0.881

These results were calculated assuming a disease causing allele with a relative risk of 1.5, a minor allele frequency of at least 0.05, that a p-value threshold of  $5 \times 10^{-7}$  is used to define power and that the study should consist of an equal number of cases and controls. The second column shows the prices that we were able to obtain for these products at the time of submission. The last line of the table shows the power that would be obtained using the 'Complete' chip using the sample size equal to that of the most powerful design.

doi:10.1371/journal.pgen.1000477.t002

# Additional Considerations for an Association Study

- Population substructure
- In addition to single marker analyses; multi-marker (haplotype) should be included whenever possible
- Genotype data quality controls
- Cross-platform issues
- Gene X Gene
- Gene X Environment
- Follow-up investigation of associated region
- Replication Replication Replication

# So you want to do a gene expression study...

- What type of tissue to use?
- How will I stabilize my RNA?
- How many genes do I want to evaluate?
- What technique should I use for data collection?
- Sample size estimates?
- Issues with multiple platforms to collect data?
- Cross sectional or longitudinal?
  - unlike DNA, gene expression can change over time and in response to endogenous and exogenous exposures

# Online Resources to Visit

- **SNP database (dbSNP)**
  - [www.ncbi.nlm.nih.gov/projects/SNP](http://www.ncbi.nlm.nih.gov/projects/SNP)
  - Catalog of polymorphisms involving one or very few nucleotides
  - Many organisms
  - Selection criteria based on MAF, heterozygosity, function class, etc
  - Population based allele frequencies

# Online Resources to Visit

- **HapMap**

- [hapmap.org](http://hapmap.org)
- Catalog of common polymorphisms
- Goes beyond data on individual polymorphisms (for example LD measures)
- Tagging SNP selection

# Online Resources to Visit

- **Cancer Genome Anatomy Project**
  - [cgap.nci.nih.gov](http://cgap.nci.nih.gov)
  - Genomic information on normal, precancerous and cancerous cells
  - RNA interference information relative to cancer-related genes
  - Direct access to tools, pathways and databases to assist with cancer genomic research

# Online Resources to Visit

- **Gene Expression Omnibus (GEO)**
  - [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)
  - repository of gene expression data submitted by the scientific community
  - Allows for data mining based on your criteria
  - Access to tools to allow you to further evaluate the data in the repository
  - Most journals require that authors deposit data into GEO

# Online Resources to Visit

- **Database of Genotypes and Phenotypes (dbGAP)**
  - [www.ncbi.nlm.nih.gov/sites/entrez?db=gap](http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap)
  - Archives and distributes data from studies that have investigated the interaction of genotypes and phenotypes

# Online Resources to Visit

<b>Exemplars of GWAS in dbGaP</b>	<b>Sample Size</b>	<b>Type of Study</b>
<b>NINDS repository Cerebrovascular Disease/Stroke Study</b>	<b>840</b>	<b>Case set</b>
<b>Framingham</b>	<b>14,276</b>	<b>Longitudinal</b>
<b>NIDDK IBD GC Crohn's disease GWAS</b>	<b>1,963</b>	<b>Case-Control</b>
<b>Study of Addiction Genetics and Environment (SAGE)</b>	<b>4,121</b>	<b>Case-Control</b>
<b>SNP Health Association Resource (SHARe) Asthma Resource Project (SHARP)</b>	<b>4,046</b>	<b>Longitudinal</b>
<b>GENEVA Diabetes Study</b>	<b>6,033</b>	<b>Nested Case-Control</b>
<b>Whole Genome Association Study of Systemic Lupus Erythematosus</b>	<b>4,651</b>	<b>Case-Control</b>
<b>IMSGC Genome Wide Association Study of Multiple Sclerosis</b>	<b>3,002</b>	<b>Parent-Offspring Trios</b>

# Online Resources to Visit

- **Catalog of published GWA studies**

- [www.genome.gov/26525384](http://www.genome.gov/26525384)
- Can search by disease/trait; gene, chromosomal region, SNP, p-value
- As of 10/06/09, the table includes 416 publications and 1863 SNPs
- Table includes information on sample size (for initial and replication sample) and platform as well as other relevant information

Hindorff et al (2009) Potential etiologic and functional Implications of genome-wide association loci for human diseases and traits PNAS 106(23):9362-9367

# Online Resources to Visit

- **Human Genome Resources Portal**
  - [www.ncbi.nlm.nih.gov/projects/genome/guide/human](http://www.ncbi.nlm.nih.gov/projects/genome/guide/human)
  - Access to many of the online resources already mentioned...and then some!

# Parting Words...

- Visit and Exploit online resources as much as possible
- Don't reinvent the wheel – move science forward
- The techniques to collect and analyze genetic/genomic data advances rapidly – keep yourself informed
- The study's purpose should drive study design – but we do need to be practical...