

ISGC Grant Development Workshop
Salt Lake City, Utah
April 3-4, 2014

*Using Extreme Phenotypes to
Identify Causal (and Protective)
Variants in Ischemic Stroke*

John W. Cole
University of Maryland

Outline

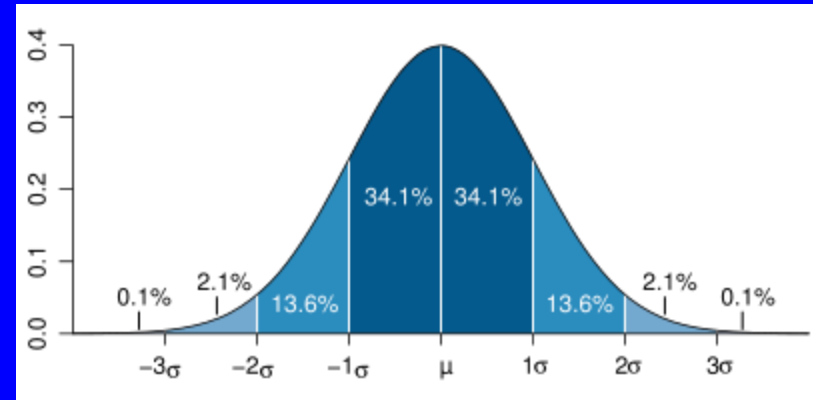
1. Background
2. Potential Study Designs

What is an Extreme Phenotype?

Outliers of distributions

Case examples

- No risk factors
- Early age-of-onset with few risk factors
- + Family History



Potential Control example

- Many risk factors (no stroke) - protective

Extreme Phenotypes and Genetics

Extreme phenotypes are enriched for genetic variation in the same genes and pathways that cause the common forms of stroke.

Extreme phenotypes are enriched for:

1. Rare variants – multiple different low frequency variants in same genes or pathways.
2. Common variants - not identified due to noise or heterogeneity within the larger sample.

Identifying Extreme Phenotypes: Appropriate case and control selection is the key

Need to consider what one is trying to identify?

Is the goal to explore for:

1. Pure stroke risk variants - stroke risk independent of risk factors
2. Stroke risk variants - mediated by the presence of a risk factor (i.e. HTN, DM, smoking, LDL, +FH, others)

Considerations

- The more homogeneous the extreme phenotype, that is with fewer confounding variables, the more likely one will be able to identify risk variants.
- Caveat: Progressively improving homogeneity - by removing confounding variables, progressively reduces sample size
 - can lead to reduced power

Key Question

Can focusing on extreme phenotypes help uncover stroke-related genes and pathways?

Proposal Hypothesis

- SiGN population outliers (i.e. extreme phenotypes) are enriched for low frequency risk variants that cluster within the same stroke susceptibility genes and pathways as within the total population (all stroke and by subtype), but are not well tagged by GWAS common variants.

Strategies to identify extreme phenotypes in stroke

1. Cases with early-onset (and no known risk factors)
2. Cases with low genetic risk for stroke based on polygenic risk scores
3. Cases with low environmental risk for stroke based on absence of stroke risk factors
4. Others??

All stroke and by subtypes

Optimal Study Design?

- Many possible.....
- Ideally, utilize all SIGN participants
 - All have GWAS data
 - Many have exome chip content
 - On average, cases with more phenotype data than controls
- To isolate individuals with extreme phenotypes develop and implement:
 1. polygenic risk score
 2. poly-phenotypic risk score

Polygenic and Poly-Phenotypic Risk Scores

For all cases and controls in SiGN calculate:

- Polygenic risk score - assess each individual's burden of the 10-20K most strongly associated SiGN GWAS SNPs.
- Poly-phenotypic risk score - based on each individual's burden of vascular risk factors
 - DM, HTN, Smoking, HLD, +FH, BMI, etc. (standardized method)

*A few potential study
designs to consider....*

Study Design #1

Stage 1

Identify SIGN cases with low poly-genotypic and poly-phenotypic risk scores (no GWAS risk and no RF risk).

In these individuals, screen SiGN exome chip content to identify low-frequency variants in specific genes and pathways - burden testing

Stage 2

Sequence the most informative subset of extreme cases and controls

Stage 3

Implicated SNPs/genes are verified in other populations

Stage 1: Discovery

Determine genes and pathways associated with stroke risk

- Identify SIGN cases with low poly-genotypic and low poly-phenotypic risk scores
 - no GWAS risk and no vascular RF risk.
- In these individuals, evaluate SiGN exome content with gene- and pathway-burden testing.
- Stratified analyses by subtype (TOAST and CCS), ethnicity, gender, and age-of-onset.

Stage 2: Verification

- Perform gene/pathway-specific sequencing (fine-mapping) on the most informative extreme cases and controls.
 - If few signals are seen in the SiGN exome data, then exome- or WG-sequencing could alternatively be performed on the extreme individuals.

Stage 3: Generalization

- Perform fine-mapping or SNP-specific genotyping within the identified genes and pathway-specific-genes in additional SiGN (or other) stroke cases-control samples.

Study Design #2

Considerations:

- The target is rare variants.
- If exome-chip is in the first step, then its coverage determines whether one will miss or identify variant(s).

Study Design #2

A few more considerations:

- GEOS - is an extreme phenotype population = early-onset stroke
- GEOS has GWAS data
- GEOS does not have exome chip data
 - Exome chip is cheap
 - Exome sequencing is costly but would be more informative.

Study Design #2

1. Using GEOS, discovery by exome-chip and in extreme GEOS cases and controls exome-sequencing
 - SiGN GWAS genotypic score will be used to help define extreme phenotypes in GEOS
2. Use SiGN exome content to further refine genes and pathways associated with risk/protection
3. Targeted sequencing of the discovery regions in a larger sample of extreme cases from within SiGN

Using SiGN GWAS data to determine pattern for low poly-genotypic risk scores - no GWAS risk

Screen GEOS GWAS samples to identify those low poly-genotypic risk scores



Further screen GEOS to identify samples with low poly-phenotypic vascular risk factor scores

Not Low



SiGN Exome
Chip Data

GEOS Exome Sequence

GEOS Exome Chip

Use GEOS and SiGN exome content to define genes and pathways associated with risk

Verify: Targeted sequencing of the discovery regions in a larger sample of extreme cases from within SiGN

Comments/Questions?

Study Design #1

1. Identify SIGN cases and controls with low GWAS-based poly-genotypic and poly-phenotypic risk scores.
 - Utilize SiGN exome chip content to identify low-frequency variants.
2. Sequence most informative subset of extreme cases and controls from stage 1.
3. Verify implicated loci examined in the other samples.

Study Design #2

1. Using GEOS, discovery by exome-chip and exome-sequencing in extreme cases and controls.
 - SiGN GWAS genotypic score used to help define extreme phenotype
2. Use SiGN exome content to further refine
3. Targeted sequencing of the discovery regions in larger samples of extreme cases from within SiGN