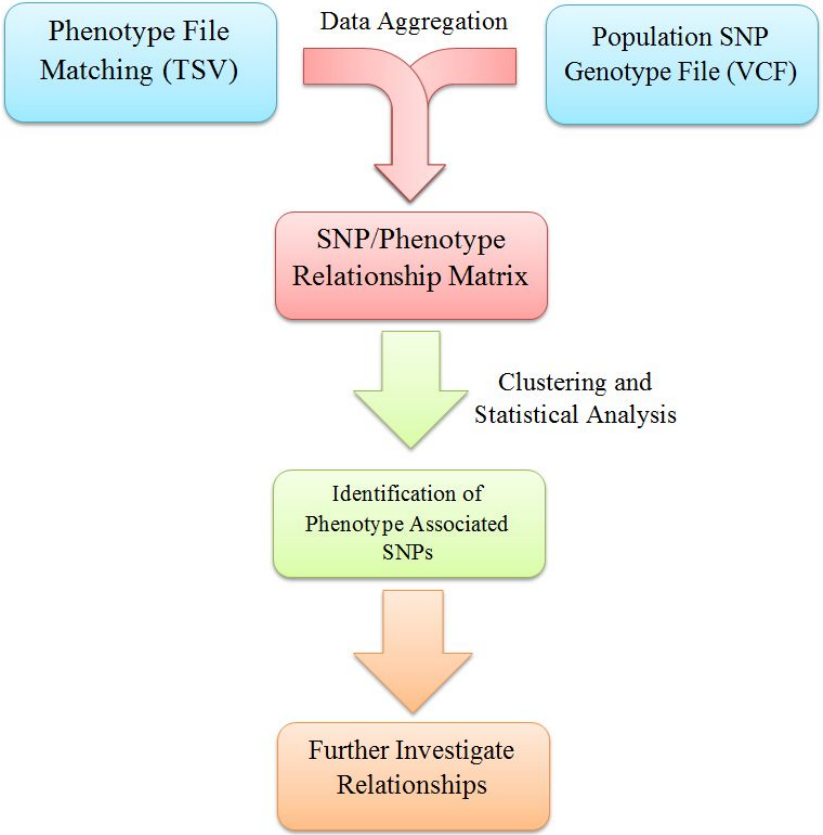
Statistical Methods for Analysis of Variants



Team Members:
Kevin Arvai
Jonathan Goodson
Brendan Joyce
Peter Merrick
Behram Radmanesh
Katie Thompson
Newton Vidal

Motivation and Introduction:

- Goal of Project: Develop pipeline to find SNP combinations related to phenotype
- Corollary project: Solve chromosomal accession interoperability issue

Pipeline Overview:

1. Preprocess data.
2. (Optional) Subset SNPs into smaller data sets for analysis.
3. Run the feasible solutions algorithm (FSA).
   -- Option 1: run FSA on subsets of SNPs.
   -- Option 2: run FSA on full data set using “fast” FSA.
4. Interpret output.

Use Case Data: GWAS on neuropathy in multiple myeloma

- Information:
  - 116 samples, all received bortezomib-dexamethasone therapy
  - 795,736 SNPs
- Genotype data: snp by snp genotyping array
- Phenotype: bortezomib-induced peripheral neuropathy (BiPN)
  - 75 controls, 41 with neuropathy grade ≥ 2
- Aim: improve prediction of susceptibility to BiPN in multiple myeloma patients
- GSE66903 (GEO database)

Simulated Data Results:

- 116 samples
- 360,000 variables
- ~37 min to analyze with one replication of algorithm

Full Output Example:

```
$solutions
  start.1 start.2 best.1  best.2  criterion swapsn checks
1  X25451 X319051 X36696 X240499 0.00426447      2    300

$table
   criterion    Var2   Var3 times warnings
1 0.00426447 X240499 X36696     1       NA

$efficiency
[1] "You did: 300  model checks compared to  64799820000  checks you would have done
with exhaustive search."
```

Application to Use Case:

- 116 samples
- 795,736 SNPs
- ?? min to analyze with one replication of algorithm

Output:

Coming soon…

Future Work:

- Explore test use case
- Linkage disequilibrium
- Other phenotypes
- SNP filtering