Introductions

Janet Sinsheimer, Eric Sobel, Hua Zhou
Biomathematics, Biostatistics, and Human Genetics
UCLA

JSM 2018

Goals

- Introduce you to the Julia language, a modern R, and a world-class vehicle for numerical computation.
  - C, C++, Fortran, and Julia are the only languages to reach PetaScale computing; and Julia the only high-level language to do so.
- Introduce you to Statistical Genetics, in particular the OpenMendel package written in Julia.

History of the Statistical Genetics Software Package, Mendel

- Designed primarily to allow linkage style gene mapping using in families. Ken Lange wrote the first version in Fortran 77 over 30 years ago. Distributed as source code, only ~4500 lines at that time.
- Current version v16 is in Fortran 95 and is >75,000 lines of code. Available for free from http://www.genetics.ucla.edu/software/
- 31 options ranging from linkage to genome-wide association studies, variance components, genetic counseling, data manipulation, diagnostics. Chief Architects: Ken Lange, Janet Sinsheimer, Eric Sobel, Hua Zhou.
- Professional user interface that combines data management of PHI, visualization tools, and statistical analysis: Mendel Enterprise. Director: Jeanette Papp

More on the Fortran Version of Mendel

- Good points
  - Very versatile, well vetted software
  - Takes many file input formats
  - Well documented. Documentation reads like a book and all answers are available in the 300 pages
- Bad points
  - Code is complex and deep: Eric, Hua and Ken are true experts on navigating the code but it's hard for students or more casual users to modify code. Not open source because of this.
  - Fortran has not kept pace with modern developments in programming, statistics, numerical analysis, computer hardware.
- All future development will be in Julia so no new options in the Fortran version.

Why Julia? (subject next few talks)

- Julia is free, easy to install, and cross-platform.
- Julia is easy to learn: a shallow learning curve.
- Julia has a clear syntax that lends itself to compact and readable code.
- Julia is fast and the code is easily parallelized.
- Julia uses an easy, modern package-management system.
- Julia has many statistical and analysis libraries - approaching R and Matlab.

What are the Goals of OpenMendel Project?

- To make the best features of Mendel easily assessable to coders.
- To allow Mendel to keep up with modern computing needs in genetics and genomics by creating a platform that can adapt to changes in technology and interests.
- To allow anyone to implement their ideas in Mendel.
- To put in place quality control and accepted software development standards and documentation.
- To create modular code and make data management easier.

Structure of OpenMendel

- There is base code and contributed packages that use the base code.
- Data entry and output uses data frames, matrices with different column types that can allow for missing data (similar to R).
- When the data sets are large, like genome-wide SNP data, can use compressed binary formats (e.g., via the SnpArray package).

Examples of Recent OpenMendel Projects at UCLA

- Variance component models for GWAS (Hua Zhou)
- Model selection in random sample GWAS (Ben Chu, Kevin Keys, and Ken Lange)
- Trait simulation (Huwenbo Shi, Ken Lange, Hua Zhou, and Eric Sobel)
- Genotype and haplotype imputation (Rory Wasiolek, Ken Lange, and Janet Sinsheimer)

MendelBase Package

- Handles data structures and keywords
- Input of data
- Pedigree computation (Elston-Stewart algorithm)
- General and Genetic utilities
- Model construction

Available Open Mendel Packages

Package Name	Purpose
Search	Numeric optimization: Non linear optimization with linear constraints
VarianceComponentModels	Numeric optimization: Fitting and testing of linear mixed effect models
SnpArrays	Data manipulation: handing of compressed biallelic SNP data
AimSelection	Data manipulation: Finds the SNPs that provide the greatest difference by population for the current sample. (AIM = ancestry informative marker).
GeneDropping	Data manipulation: genotype simulations with family data

Available Open Mendel Packages

Package Name	**Purpose**
EstimateFrequencies	Estimation: of allele frequencies of markers using pedigree data.
Kinship	Estimation: of the degree of relatedness between two individuals
GWAS	Inference: Genome-wide association tests
GameteCompetition	Inference: Genetic Association testing in pedigree data
TwoPointLinkage	Inference: of linkage by estimating and testing recombination fractions
LocationScores	Inference: multi-marker version of Two point linkage, estimates and tests putative gene location relative to markers
GeneticCounseling	Estimation: risk calculation for Mendelian disease

Input files types

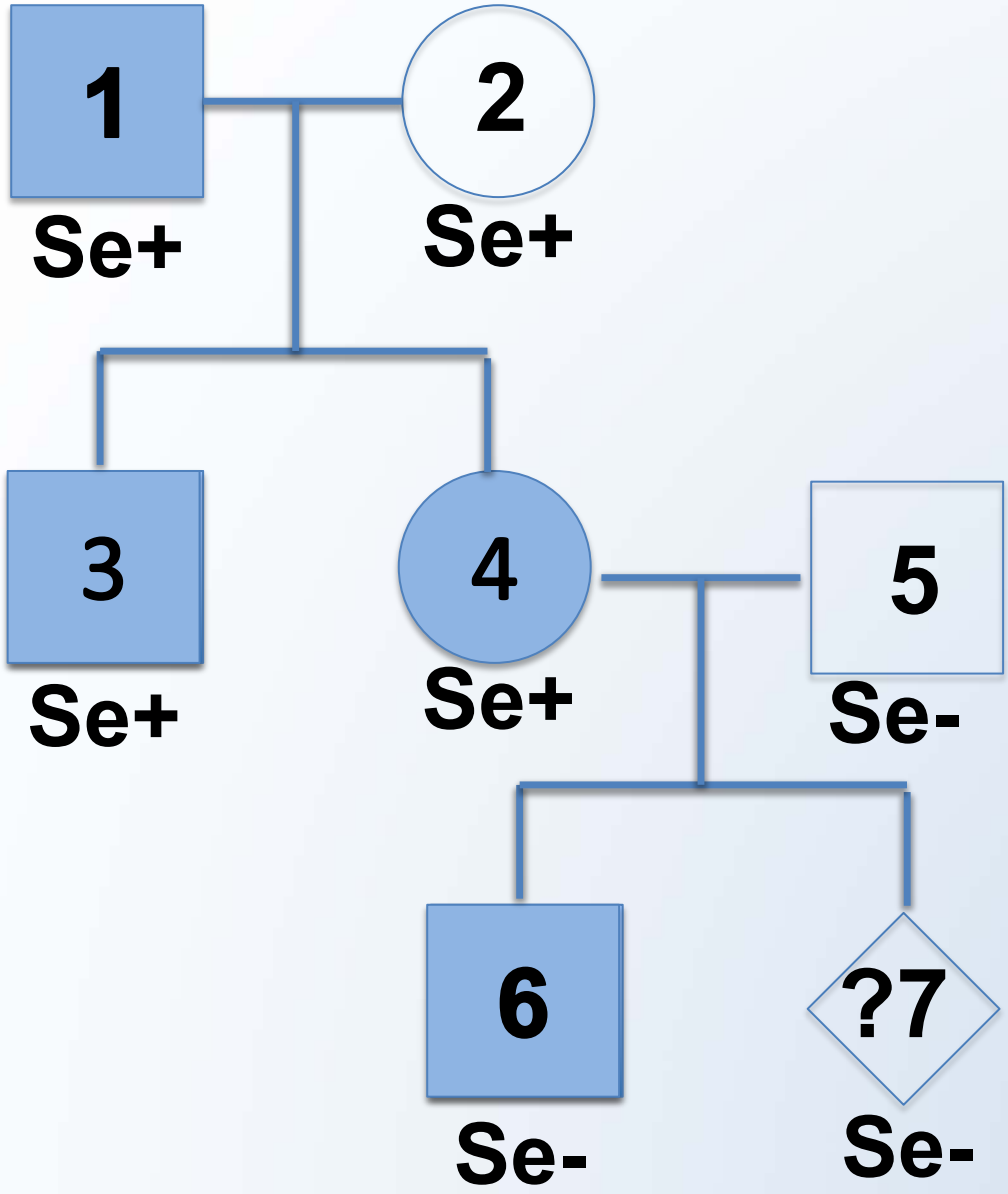
- Control File: Specifies data input and output files and any optional parameters (*keywords*) for the analysis.
- Locus File: Names and describes the genetic loci (gene locations, allele names, allele frequency)
- Pedigree File: Gives information about the individuals, such as name, sex, family structure, and ancestry.
- Phenotype File: Lists the available phenotypes.
- SNP Definition File: Defines the SNPs: SNP name, chromosome, position, allele names, allele frequencies.
- SNP Data File: Holds the SNP genotypes. Must be a standard binary PLINK BED file in SNP major format. SNP data file require a SNP definition file.

Genetic Counseling Example:

What is the probability that child 7 in the following pedigree will be affected with the rare autosomal dominant disease myotonic dystrophy given the pedigree and loosely linked dominant marker Se (secretor protein presence/absence)?

$$P(7affected \mid 7Se- \& family) = \frac{L(7affected, Se- \& family)}{L(7unknown, Se- \& family)}$$

Pedigree



Pedigree as a data frame

```
Pedigree,Person,Mother,Father,Sex,MD,Se
Top,1,,,male,Affected,+
Top,2,,,female,Normal,+
Top,5,,,male,Normal,-
Top,3,2,1,male,Affected,+
Top,4,2,1,female,Affected,+
Top,6,4,5,male,Affected,-
Top,7,4,5,male,Affected,-
Bottom,1,,,male,Affected,+
Bottom,2,,,female,Normal,+
Bottom,5,,,male,Normal,-
Bottom,3,2,1,male,Affected,+
Bottom,4,2,1,female,Affected,+
Bottom,6,4,5,male,Affected,-
Bottom,7,4,5,male,,-
```

Control File

```
#
# Input and Output files.
#
locus_file = genetic counseling 2 LocusFrame.txt
pedigree_file = genetic counseling 2 PedigreeFrame.txt
phenotype_file = genetic counseling 2 PhenotypeFrame.txt
output_file = genetic counseling 2 Output.txt
#
#  Analysis parameters for Genetic Counseling option.
#
```

Locus File

```
Locus,Allele,Chromosome,FemaleMorgans,European
MD, +,    Autosome, 0.0, 1.0e-5
MD, -,    Autosome, 0.0, 0.99999
Se, +,    Autosome, 0.08718, 0.52
Se, -,    Autosome, 0.08718, 0.48
```

Phenotype File

```
Locus,Phenotype,Genotypes
MD,Affected,"+/+,+/-"
MD,Normal,"-/-"
Se,+,"+/+,+/-"
Se,-,"-/-"
```

Result in Output File

The risk = 0.83986