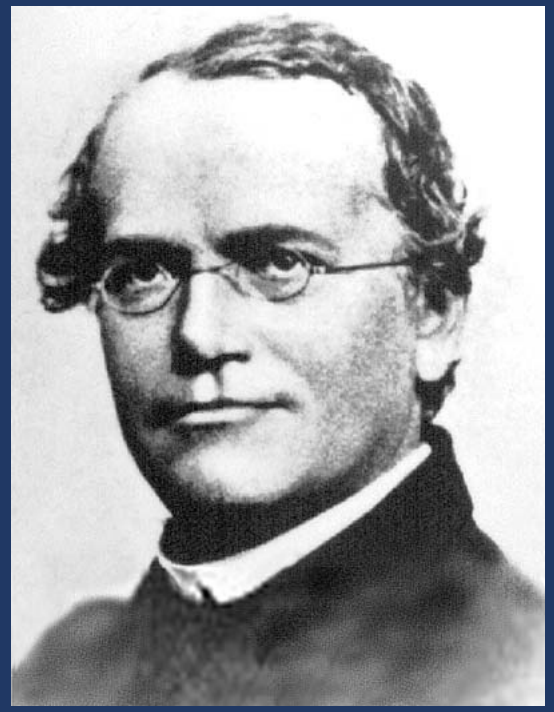
GWAS for Ordinal Phenotypes with OrdinalGWAS.jl



OpenMendel Workshop

ASHG Annual Meeting 2020

Chris German

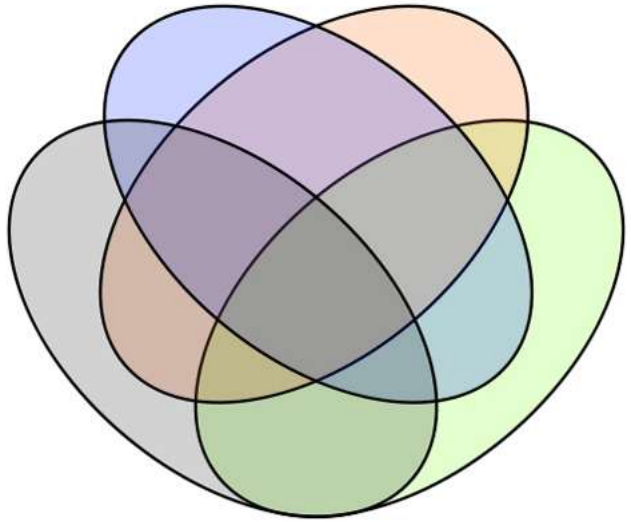
Motivation

Ordinal phenotypes, traits that take ordered, discrete values naturally occur in many health related studies

- Common in complex diseases where a binary status may not be suitable (stages of cancer)
- Electronic health record phenotyping algorithms often use many variables and have uncertainty when it comes to binary labeling
- Examples of ordinal outcomes:
  - Disease severity/progression (*undiagnosed, pre-disease, mild, moderate, severe*)
  - Disease likelihood (*unlikely, possible, probable, certain*)

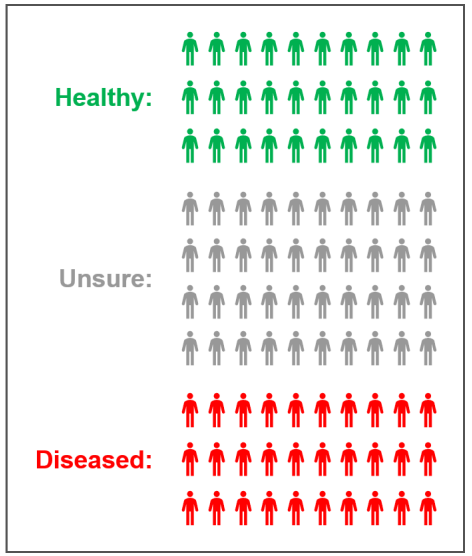
How to conduct GWAS with ordinal traits?

1. Dichotomize the ordinal trait for case/control GWAS



Run sets of case/control GWAS and look at overlap

hard to interpret results



Exclude individuals or make arbitrary cutpoints

lose power

How to conduct GWAS with ordinal traits?

2. Treat ordinal trait as linear

- Violates assumptions of linear regression - is the distance between outcomes equal? *control* and *pre-disease* same as *mild* and *moderate*?

How to conduct GWAS with ordinal traits?

3. *Use a model **for** ordinal data*

OrdinalGWAS.jl uses an ordered multinomial model:

- Assume trait Y takes J ordinal values
- Cumulative probabilities are linked to covariates $\mathbf{x}_i$ and SNPs $\mathbf{g}_i$ via

$$\text{logit}[\mathbb{P}(Y_i \leq j)] = \theta_j - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{g}_i^T \boldsymbol{\beta}_\mathbf{g}, \quad j = 1, \ldots, J-1,$$

- Intercepts $\theta_1 \leq \cdots \leq \theta_{J-1}$ enforces order between categories, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_\mathbf{g}$ are the regression coefficients of covariates and SNP(s) repspectively

Implementation

- For GWAS, we develop efficient score test (requires only null model to be fit) for testing:
  - Single SNPs
  - SNP sets (rare variants)
  - Gene-by-environment interactions
- Each test scales *linearly* with sample size
- Recommended practice:
  - Perform score test for genome-wide scan
  - Re-test the most significant ones using the likelihood ratio test (slower but higher power)
- Hypertension GWAS on a UK Biobank sample (~185,000 individuals and 460,000 SNPs) took 181 minutes on a standard PC

Usage

- **OrdinalGWAS.jl** currently works with PLINK and VCF Files
- Straightforward usage, specify the null model formula, covariate file, and genetic data file basename (no extension)

```
using OrdinalGWAS
ordinalgwas(@formula(hypertension ~ sex), "data/covariate.txt", "data/hapmap3")
```

```
StatsModels.TableRegressionModel{OrdinalMultinomialModel{Int64,Float64,LogitLink},Array{Float64,2}}

hypertension ~ sex

Coefficients:
──────────────────────────────────────────────────────────
               Estimate  Std.Error   t value  Pr(>|t|)
──────────────────────────────────────────────────────────
intercept1|2  -1.48564    0.358891  -4.13952    <1e-4
intercept2|3  -0.569479   0.341044  -1.66981    0.0959
intercept3|4   0.429815   0.339642   1.26549    0.2066
sex            0.424656   0.213914   1.98517    0.0480
──────────────────────────────────────────────────────────
```

Usage

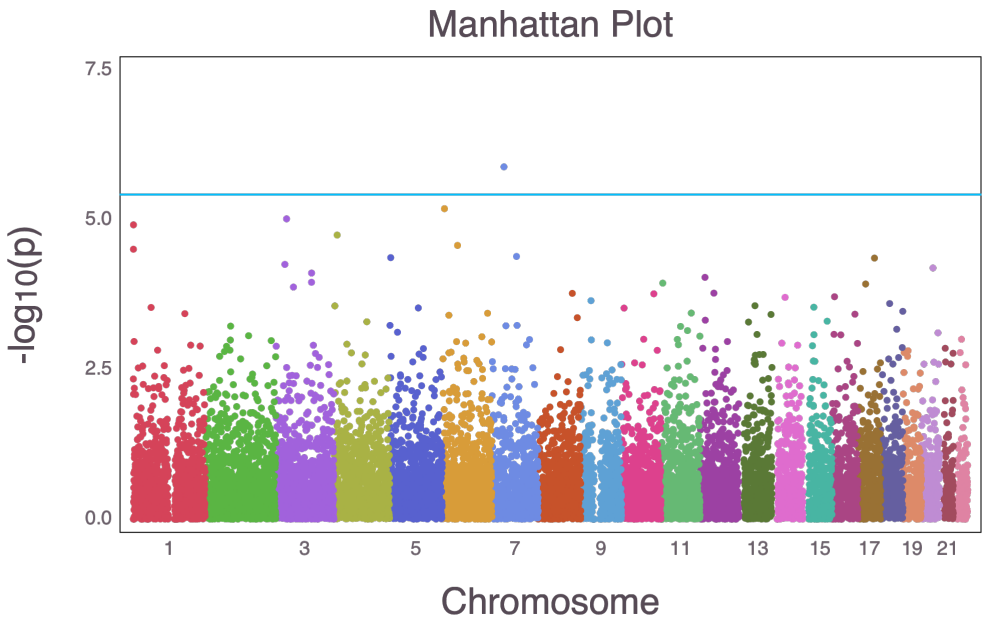
- **OrdinalGWAS.jl** produces an output file of each SNP's results

```
using CSV
results = CSV.read("ordinalgwas.pval.txt")
first(sort(results, :pval), 8) # sort and display top 8 p-values
```

8 rows × 6 columns

	chr	pos	snpid	maf	hwepval	pval
	Int64	Int64	String	Float64	Float64	Float64
1	7	41152376	rs28880	0.337963	0.805237	1.30801e-6
2	6	7574576	rs1885466	0.177469	0.762069	6.53672e-6
3	3	36821790	rs4678553	0.234568	0.109467	9.66474e-6
4	1	1168108	rs11260566	0.191589	0.128568	1.21687e-5
5	4	11017683	rs16881446	0.275542	0.894275	1.80275e-5
6	23	121048059	rs1937165	0.43808	3.95961e-16	2.09895e-5
7	6	52474721	rs2073183	0.182663	0.507777	2.68445e-5
8	1	967643	rs2710875	0.324074	4.07625e-7	3.10828e-5

```
using MendelPlots
manhattan(results)
```



Usage

- Top SNPs can be selected, and rerun using LRT in OrdinalGWAS.jl to get effect size estimates

	chr	pos	snpid	maf	hwepval	effect	pval
	Int64	Int64	String	Float64	Float64	Float64	Float64
1	1	967643	rs2710875	0.324074	4.07625e-7	-0.648856	1.80505e-5
2	1	1168108	rs11260566	0.191589	0.128568	-0.915723	5.87338e-6
3	3	36821790	rs4678553	0.234568	0.109467	0.742495	1.13038e-5
4	4	11017683	rs16881446	0.275542	0.894275	-0.787058	1.11054e-5
5	5	3739190	rs12521166	0.0679012	0.186136	1.14689	4.78129e-5
6	6	7574576	rs1885466	0.177469	0.762069	0.875062	7.27235e-6
7	6	52474721	rs2073183	0.182663	0.507777	0.779079	5.06939e-5
8	7	41152376	rs28880	0.337963	0.805237	-0.814634	9.18013e-7
9	7	84223996	rs4128623	0.0787037	0.0218347	1.00222	6.5879e-5
10	23	121048059	rs1937165	0.43808	3.95961e-16	0.539231	1.97546e-5

Other Features

We go over the following features in the binder tutorial:

- Using VCF Files
- Filtering:
  - Restricting to some individuals
  - Restricting to some SNPs
- Link functions
- GxE testing
- SNP-set testing

Methods and applications of OrdinalGWAS.jl are detailed in the publication:

- German CA, Sinsheimer JS, Klimentidis YC, Zhou H, Zhou JJ. (2020) Ordered multinomial regression for genetic association analysis of ordinal phenotypes at Biobank scale. Genetic Epidemiology. 44:248-260. https://doi.org/10.1002/gepi.22276