Copula Modeling of Dependent Traits in Rare Variant Analysis

Yildiz Yilmaz

Department of Mathematics and Statistics
&
Discipline of Genetics, Faculty of Medicine
Memorial University of Newfoundland

August 9, 2018

Complex traits

- Many genetic association studies have been conducted to identify genetic variants associated with complex traits.
- However, much of the heritable variation in complex traits is still unexplained.
- There are many genetic and environmental factors that affect complex traits.
- Genetic factors may include some common ($MAF \geq 0.05$), low-frequency ($0.01 \leq MAF < 0.05$), and rare ($MAF < 0.01$) genetic variants.

Rare variant analysis

- Possible approaches to detect variants with small effects or rare variants:
  - Increase the sample size
  - Improve the study design: Reduce the phenotypic and genetic heterogeneity - select a more homogenous subgroup of individuals
  - Use appropriate methods of analysis
- Single-marker tests are often the method of choice for the analysis of common or low-frequency genetic variants.
- In population-based studies, single-variant analysis of rare variants may yield low power if the effect of the causal variant is not large.
- Thus, recent studies have focused on developing multi-marker rare variant association tests to identify causal genomic regions.

Multi-marker tests

- Multi-marker tests aggregate association signals across multiple rare variants in a genomic region.
- For population-based studies, some multi-marker tests were proposed.
- Lee et al. (2014) give a nice summary of different types of tests:
  - Different classes of methods including burden tests (e.g., Li and Leal, 2008; Morris and Zeggini, 2010), variance-component tests (e.g., Wu et al., 2011), combination of burden and variance-component tests (Derkach et al., 2013; Lee et al., 2012).
- Power of these tests depends on the proportion, effect sizes and directions of the effects of causal (in fact, associated) variants in a given region.

Single- versus multi-marker tests

- The aim of the multi-marker tests is to identify genomic regions associated with the trait.
- Multi-marker tests are testing
  - whether a given combination of variants in a given gene is associated with the trait (burden-type tests)

  or
  - whether any of the variants in a given gene is associated with the trait (variance-component-type tests).
- Single-marker tests are testing whether a given variant is associated with the trait.

Single- versus multi-marker tests

- To fairly compare the performance of these two types of tests, we need to compare them in their power to identify the same causal genetic locus (e.g., a gene).
- Thus, for single marker tests, we test whether any of the rare variants within the gene shows a significant association with the trait while accounting for multiple testing.
- We compared a single-marker test with some multi-marker tests (a burden test, SKAT, SKAT-O) for testing the same hypothesis in rare variant association studies of quantitative traits (Konigorski et al., 2017).

Single- versus multi-marker tests

- We considered a linear regression model of a normally distributed quantitative trait.
- We observed that the least square estimation method and the t-test statistic have valid properties even when investigating singletons and doubletons.
- The single-marker test has larger or equal power compared to multi-marker tests as long as there is not a large number of causal variants in a region all with small effect sizes (Konigorski et al., 2017).
- The single-marker test and the multi-marker tests are all sensitive to misspecification of the error distribution.
- The distribution assumptions need to be assessed before conducting the association tests.

Joint modeling of multiple traits

- Power of the single-marker tests could be improved by incorporating additional information through modeling multiple traits.
- Suppose there are bivariate traits (Y_1, Y_2).
- Well-known joint modeling approaches are
  - Conditional analysis of traits: It consists of modeling the marginal distribution of Y_1 given covariates and modeling the conditional distribution of Y_2 given Y_1 and covariates through some regression modeling approaches.
  - Models with random effects: A bivariate random effect model assumes that Y_1 and Y_2 are independent given an unobserved random variable and covariates.
  - Marginal approach: The joint distribution of Y_1 and Y_2 is modeled directly. The marginal distributions are usually modeled seperately from the dependency structure.

Proposed methods

Some different joint modeling approaches and association tests have been proposed for genetic association studies:

- Yang and Wang (2012) and Zhu et al. (2015) discuss some joint modeling approaches and methods for joint association analysis of multiple phenotypes: modeling with random effects, variable reduction methods, combining test statistics from univariate analyses.
- MultiPhen (O'Reilly et al., 2012): Models the association between linear combinations of phenotypes and the genotypes at each variant and identifies the linear combination of the phenotypes most associated with the variant.
- MURAT (Multivariate Rare-Variant Association Test; Sun et al., 2016): A region-based rare variant association test obtained under a multivariate model of phenotypes with random variant effects. It reduces to SKAT when there is one phenotype.

Proposed methods

- aSPU, aSPUset, aSPUset-Score tests (Kim et al., 2016):
  - Fit the multivariate generalized linear model of traits conditional on a single variant (aSPU) or multiple variants (aSPUset, aSPUset-Score) using generalized estimating equations method.
  - Obtain the most powerful test statistic among different combinations of power of score test statistics over all traits (and variants).
  - aSPUset test includes some different other well-known multi-marker rare variant tests.

Comparison of modeling approaches

- Conditional modeling and random effect modeling may not give a simple form for the marginal models of phenotypes.
- Under the random effect modeling, the assumed distribution for the random effect cannot be assessed.
- Under the marginal approach, the marginal models have easily interpretable forms because they allow us to specify them according to the modeling needs.
- Copula modeling is a marginal approach.
- Copulas are functions used to construct a joint distribution function (or survival function) by combining marginal distributions with a dependence structure.

Copula modeling

- Let $g_1, g_2, ..., g_M$ denote the causal genetic variants and $\mathbf{z}$ denote the vector of other factors affecting Y_1 and/or Y_2.
- Suppose the marginal distributions of Y_1 and Y_2 conditional on covariates $\mathbf{x} = (\mathbf{z}, g_1, g_2, ..., g_M)$ are denoted by $F_1(y_1|\mathbf{x})$ and $F_2(y_2|\mathbf{x})$.
- Marginal distributions can come from any distribution family and can be different.
- The joint distribution of Y_1 and Y_2 conditional on the covariate vector $\mathbf{x}$ is constructed by combining the marginal distributions $F_1(.|\mathbf{x})$ and $F_2(.|\mathbf{x})$ using a copula function C_ψ with dependence parameter vector ψ:

$$F(y_1, y_2|\mathbf{x}) = C_\psi \left(F_1(y_1|\mathbf{x}), F_2(y_2|\mathbf{x})\right)$$

Copula modeling

- If F_1 and F_2 are continuous, there exists a unique copula function constructing the bivariate distribution function (Sklar, 1959).
- Copulas allow investigation of the marginal effects separately from the dependence structure between phenotypes since the measures of dependence do not appear in the marginal distributions.
- This allows us
  - to estimate and test the effect of a genetic variant on each trait, and
  - to identify pleiotropic variants which explain the dependence between the phenotypes (Konigorski et al., 2014).

Copula modeling

- A copula function which allows to model a variety of dependence structures could be considered.
- For example, we use the two-parameter copula function

$$C_{\phi,\theta}(u_1, u_2) = \left[\left((u_1^{-\phi} - 1)^{\theta} + (u_2^{-\phi} - 1)^{\theta}\right)^{1/\theta} + 1\right]^{-1/\phi},$$

  which allows a flexible modeling and contains the Clayton (when $\theta = 1$), the Gumbel-Hougaard (when $\phi \to 0$), and the independent (when $\theta = 1, \phi \to 0$) copula (Joe, 1997).
- It is a member of the Archimedean copula family which contains some bivariate random effect models (Oakes, 1989).

Copula modeling

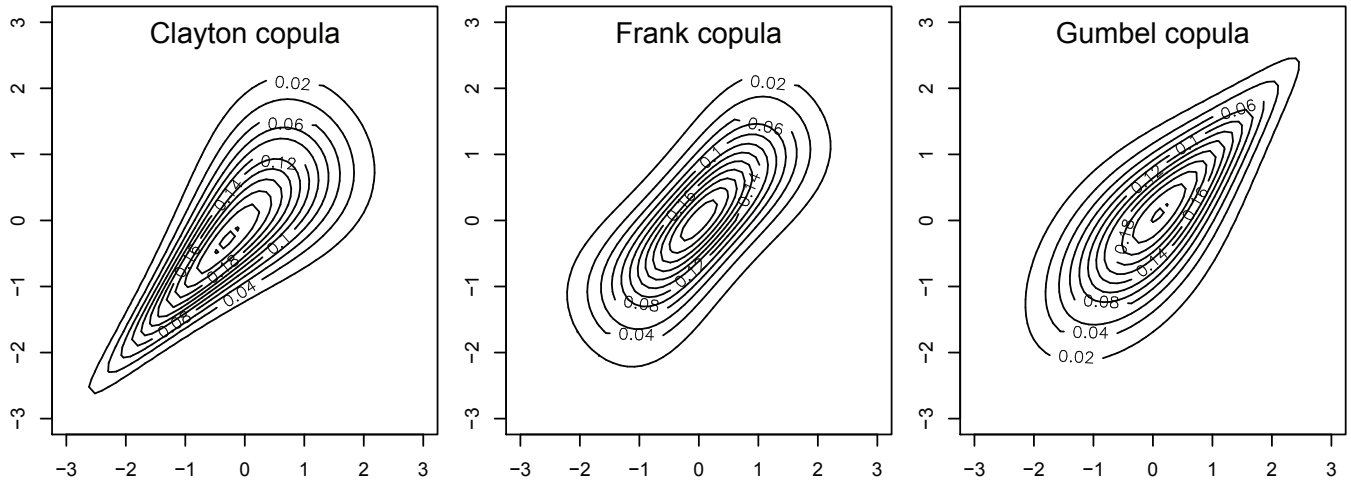

Figure 1: Density contour plots of bivariate distributions using Clayton, Frank, and Gumbel-Hougaard copulas when Kendall's $\tau = 0.5$ with standard normal margins.

- The Clayton copula has lower tail dependence but no upper tail dependence (Clayton, 1978).
- The Gumbel-Hougaard copula has upper tail dependence but no lower tail dependence (Gumbel, 1960).

Marginal models of phenotypes

- Suppose the marginal models are in the form of

$$Y_1 = \alpha_0 + \alpha_{\mathbf{1}}\mathbf{z} + \sum_{j=1}^{M} \alpha_{2j} g_j + \epsilon_1$$

$$Y_2 = \beta_0 + \beta_{\mathbf{1}}\mathbf{z} + \sum_{j=1}^{M} \beta_{2j} g_j + \epsilon_2.$$

- Distributions of ϵ_1 and ϵ_2 could come from any distribution family.
- In our simulation study, we assume that ϵ_1 and ϵ_2 come from Normal distributions with mean 0 and constant variances.

C-JAMP: Copula-based Joint Analysis of Multiple Phenotypes

- In single-marker analysis, we consider the marginal models

$$Y_1 = \alpha_0^* + \alpha_{\mathbf{1}}^* \mathbf{z_1} + \alpha_{2j} g_j + \epsilon_1$$
$$Y_2 = \beta_0^* + \beta_{\mathbf{1}}^* \mathbf{z_1} + \beta_{2j} g_j + \epsilon_2.$$

- For the genetic variant g_j, the null hypothesis in interest could be

$$H_0 : \alpha_{2j} = 0 \qquad \text{or} \qquad H_0 : \beta_{2j} = 0.$$

- The bivariate distribution of Y_1 and Y_2 given $\mathbf{z_1}$ and g_j is modeled by using a copula function

$$F(y_1, y_2 | \mathbf{z_1}, g_j) = C_\psi \left(F_1(y_1 | \mathbf{z_1}, g_j), F_2(y_2 | \mathbf{z_1}, g_j) \right).$$

- Maximum likelihood estimation is used to fit the model.
- Wald test statistic is used to test the null hypothesis.

Simulation Study - Data Generation

Construct $N = 10,000$ datasets for power comparison and $N = 100,000$ datasets for assessing type I error, each of sample size $n = 1,000$:

- Genetic data generation was similar to Lee et al. (2012).
- Generate traits Y_1 and Y_2 given the covariates $\mathbf{x} = (\mathbf{z}, g_1, ..., g_M)^T$ from the Clayton copula model with Gaussian marginal distributions.
- Weak (Kendall's tau, $\tau = 0.2$), moderate ($\tau = 0.5$) and strong ($\tau = 0.8$) dependences between the adjusted traits for covariates were considered.
- Causal SNVs have MAF ≤ 0.03.
- For effects of causal SNVs, used the scenarios in Lee et al. (2012) with $10\%, 20\%, 50\%$ causal SNVs (among SNVs having MAF ≤ 0.03), effect sizes are inversely proportional to their MAFs, and with 100%, 80%, or 50% of effects in the same direction.

Simulation results - Evaluation of asymptotic properties

- We assessed the asymptotic properties of maximum likelihood estimation under single marker analysis.
- When the *MAC* of a variant is not very low, asymptotic properties of the maximum likelihood estimation are valid.
- When the *MAC* is low and the dependence between traits is moderate or strong, asymptotic properties of the maximum likelihood estimation do not hold.
- For such variants,
  - the p-values for the Wald test can be obtained by conducting a parametric bootstrap under the estimated null model

  or

  - the distribution of the Wald test can be approximated by conducting a Monte Carlo simulation study under the estimated null model.

Simulation results - Type I error

- We test the null hypothesis that the gene is not associated with the trait Y_2.
- We consider the scenarios where
  - $\alpha_{2j} = \beta_{2j} = 0$ for all js in the gene.
  - $\alpha_{2j} \neq 0$ for some j in the gene but $\beta_{2j} = 0$ for all j.
- The empirical type I errors of C-JAMP are generally close to the nominal levels considered.
- However, when there is strong dependence between traits and the gene affects Y_1, the type I error is slightly inflated.
- When the copula model is misspecified, empirical type I error rates remain close to the nominal value.

Simulation results - Type I error

- We compared the performance of C-JAMP with MultiPhen, MURAT, aSPU, aSPUset, aSPUset-Score.
- MultiPhen, MURAT, and aSPU yielded inflated type I error rates under the assumed copula model with Gaussian marginal distributions.
- aSPUset test yields valid type I error rates and aSPUset-Score test has slightly inflated type I error rate.

Simulation results - Power

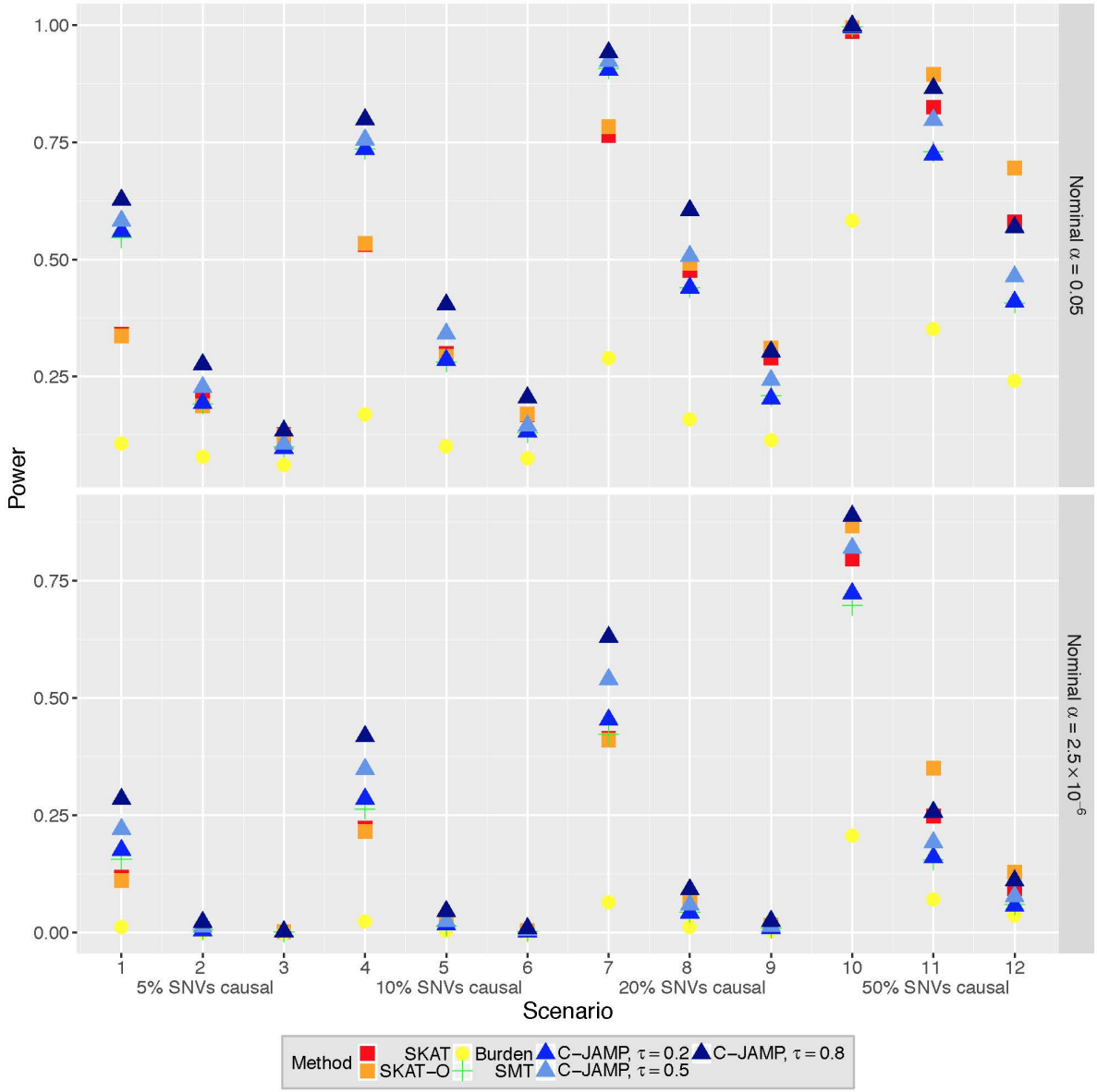

Figure 2: Empirical power estimates of C-JAMP versus the univariate SMT and MMTs

Power comparison of C-JAMP with the univariate SMT and MMTs

- Comparison to the univariate SMT, C-JAMP yields higher power when there is dependence between traits.
- As the dependence level between traits increases, power of C-JAMP increases.
- C-JAMP is more powerful than univariate MMTs except when there is a large number of causal variants all with small effect sizes.
- The power of C-JAMP is not affected by the direction of the variant effects.

Simulation results - Power

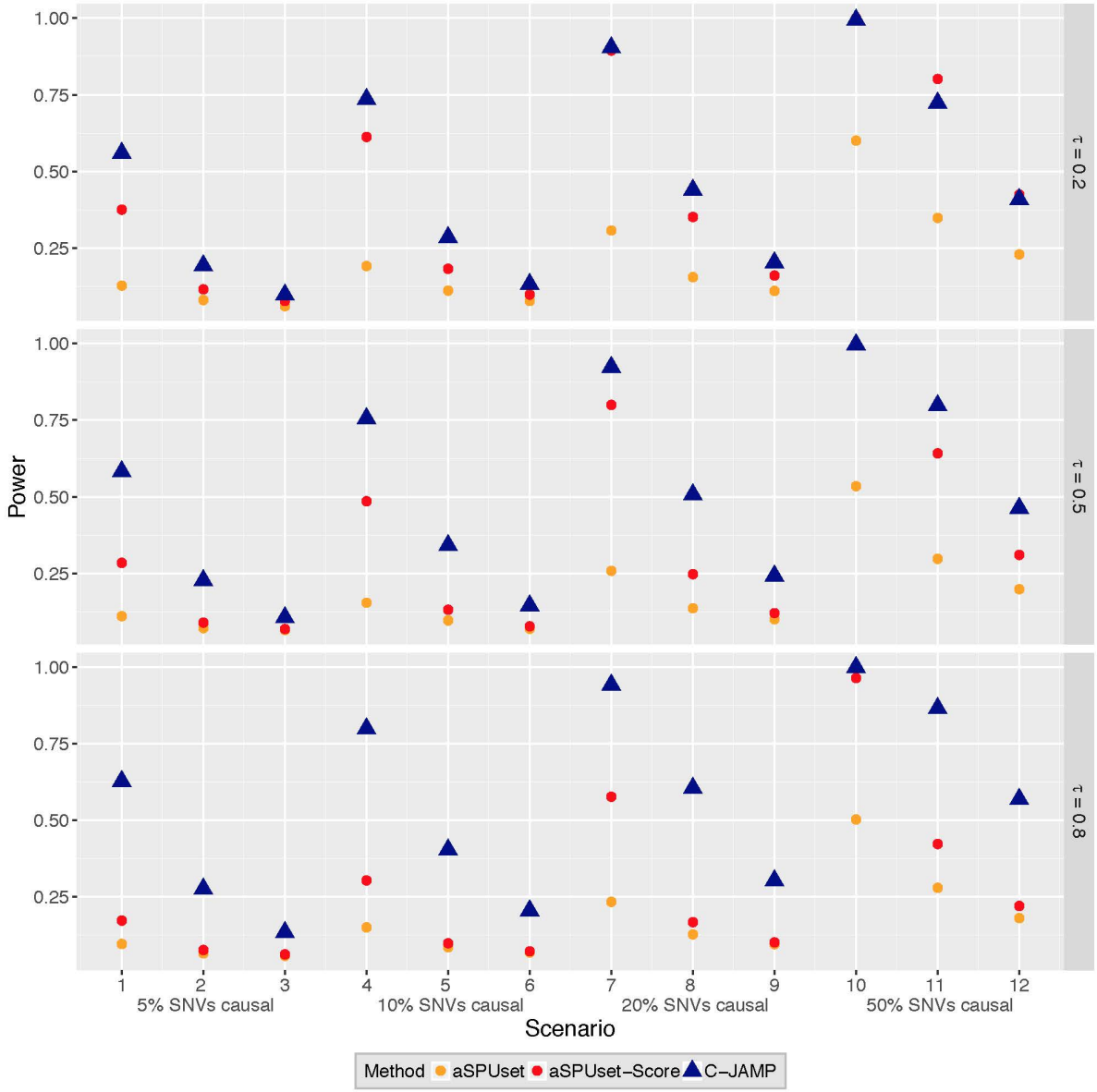

Figure 3: Empirical power estimates of C-JAMP versus multivariate MMTs

Power comparison of C-JAMP with multivariate MMTs

- Power of aSPUset-Score is always higher than that of aSPUset.
- Power of aSPUset and aSPUset-Score is very sentitive to the misspecification of dependence structure as their power decreases when the dependence level increases.
- C-JAMP yields more powerful tests except when the dependence level is low and there is a large number of causal variants all with small effect sizes.

Extension and application areas of C-JAMP

- The approach could easily be extended to the analysis of multivariate time-to-event phenotypes (Yilmaz and Lawless, 2011).
- Semiparametric estimation could be performed to reduce the marginal distribution assumptions for phenotypes (Yilmaz and Lawless, 2011).
- Other test statistics including likelihood ratio or score test statistic could be used to test the genetic association.
- The approach could be applied for the analysis of family data.
- Multi-marker tests could be obtained under copula modeling (Lakhal-Chaieb et al., 2016).

Acknowledgements

- Joint work with
  - Stefan Konigorski, Postdoctoral researcher at Max Delbruck Center (MDC) for Molecular Medicine, Berlin, Germany
  - Tobias Pischon, Molecular Epidemiology Research Group, Max Delbruck Center (MDC) for Molecular Medicine, Berlin, Germany

References

- Derkach et al. (2013). *Genet Epidemiol* 37: 110-121.
- Kim et al. (2016). *Genetics* 203: 715-731.
- Konigorski et al. (2014). *BMC Proc* 8(Suppl 1): S72.
- Konigorski et al. (2017). *PLoS One* 12(5): e0178504.
- Lakhal-Chaieb et al. (2016). *Statist Med* 35: 905-921.
- Lee et al. (2012). *Biostatistics* 13: 762-775.
- Lee et al. (2014). *Am J Hum Genet* 95: 5-23.
- Li and Leal (2008). *Am J Hum Genet* 83: 311-321.
- Morris and Zeggini (2010). *Genet Epidemiol* 34: 188-193
- O'Reilly et al. (2012). *PLoS One* 7(5): e34861.
- Sklar (1959). *Publ Inst Statist Univ Paris* 8: 229-231.
- Sun et al. (2016). *Eur J Hum Genet* 24: 1344-1351.
- Yang and Wang (2012). *J Probab Stat* 2012: 652569.
- Wu et al. (2011). *Am J Hum Genet* 89: 82-93.
- Yilmaz and Lawless (2011). *Lifetime Data Anal* 17: 386-408.
- Zhu et al. (2015). *Hum Hered* 80: 144-152.