Scripts for the paper:
Vattikuti S, Lee JL, Chang CC, Hsu SDH, Chow CC. Application of compressed sensing to genome wide association studies and genomic selection. GigaScience (in review)

Please cite the above if you use this approach for any published work.

The paper shows the applicability of compressed sensing (CS) theory to genome-wide association studies (GWAS), where the purpose is to find trait-associated tagging markers (genetic variants). CS is not a method *per se* but may be considered a general theory of regression that takes into account model complexity (sparsity). The theory is still valid in the classical regression domain of $m > n$ (where m is sample size, subjects and n are the number of measures, genetic variants) but establishes conditions for when full selection of nonzero coefficients is still possible when $m < n$. (Note the substitution of labels here compared to the paper.) A major characteristic of this is the existence of a phase transition between poor and full selection, where full selection is recovery of a solution vector of only true positives [1–5]. A true positive may be either a causal genetic variant or proxy to (in LD with) a causal variant carrying a true signal [6].

In particular, CS theory provides a mathematical justification for the use of L1-penalized regression to recover sparse vectors of coefficients and highlights the difference between *selection* of the markers with nonzero coefficients and *fitting* the precise coefficient values. In this regard it is well-known in the CS literature that L1-algorithms are close to minimax optimal and hence the phase transition discovered by these algorithms is robust to the distribution of coefficient magnitudes [6–9]. While robust in this sense, the upper bound may be improved by more informed algorithms for some distributions and signal-to-noise ratios [9]; though, this has not been shown for GWAS.

To illustrate and test these properties, we used a lasso algorithm where the penalization parameter is calculated based on an estimate of the signal-to-noise ratio given by the genomic-relatedness method (see our **MVMLE** repository for details and scripts) [10–13]. We also introduced a measure for discovering the phase transition that is novel to the CS literature and may be used for applications beyond GWAS. In this repository we include:

1 The least absolute shrinkage and selection operator (lasso) algorithm

This is an implementation of the pathwise coordinate optimization method by Friedman et al. 2007 with a warm start as suggested by Friedman et al. 2010 [14, 15]. This involves cycling down over a range of penalization parameters, λ. The upper value, λ_{max}, is the maximum coefficient magnitude given by $\mathbf{A}'\mathbf{y}$. In other words the lasso starts at the minimum point where no coefficients are selected. Pathwise coordinate optimization proceeds for each λ decreasing logarithmically until the lowest value, λ_{min}. This value is given by the maximum spurious coefficient expected due to noise (see Methods in ref. [6] and ref. [5] for details). It is employed by calling the function *lasso_sv.m* as

xhat=lasso_sv(G,miss_value,y,h2)

or

[xhat,xhatFE]=lasso_sv(G,miss_value,y,h2,FE,FEflag)

where $\mathbf{G}$ is the genotype matrix (aka *measurement matrix*) in $\mathbb{R}^{m\times n}$ with m subject rows and n SNP columns, $\mathbf{y}$ is an $\mathbb{R}^{m\times 1}$ vector of standard normalized phenotype values (continuous variable), and $h2$ is the estimated narrow-sense heritability and can be estimated using the mixed linear model from the **MVMLE** repository. If the $h2$ argument is unknown then an initial point maybe the harshest λ_{min} assuming $h^2 = 0$. The second form of implementation shows the optional arguments and return values. This form allows for fixed effect covariates. Note that it has not been thoroughly tested. The FE and $FEflag$ arguments must come as a pair. FE is a matrix of fixed effects in $\mathbb{R}^{m\times f}$ where f are the number of fixed effects. It is assumed to have no missing values. The $FEflag \in \{$'keep','select'$\}$ tells lasso to either always keep these parameters or to include them in the selection process, respectively. All data to *lasso_sv* are standard normal-

ized within the function.

A similar function has been ported to Plink 2 for the analysis of large genomic datasets.

2 Selection measures

Selection quality may be measured in a number of ways and the sample script noted below includes measures from ref. [6]. These include the normalized coefficient error (NE), positive predictive value (PPV), false positive rate (FPR), and median P-value ($\mu_{P-value}$).

The normalized coefficient error (NE) is

$$\frac{\|x - \hat{x}\|_{L_2}}{\|x\|_{L_2}}$$

The false positive rate (FPR) is the fraction of true zero-valued coefficients that are falsely identified as nonzero. The positive predictive value (PPV) is the number of correctly selected true nonzeros divided by the total number of nonzeros returned by the selection algorithm. $1 - PPV$ equals the false discovery rate (FDR). To include proxies for the PPV and FPR calculations substitute a boolean vector for $\mathbf{x}$ where one indicates an acceptable true positive.

The median of the P-values for the set of putative nonzeros ($\mu_{P-value}$) is obtained by: 1) regressing the phenotype on each of the L_1-selected markers in turn, 2) estimating each P-value as the standard two-tailed probability from the t test of the null hypothesis that a univariate regression coefficient is equal to zero, and 3) taking the median over the independent tests.

3 Example script and data

Included are two example scripts for running the sample size scan as in ref. [6]. The scripts reference functions and data in our **GD** repository. To

run these examples download this to a separate folder and add the local **GD** repository to the MATLAB path.

The file *example1_samplesize_scan.m* simulated independent Binomal random variables (sampled twice) using the chromosome 22 minor allele frequencies; whereas, the file *example2_samplesize_scan.m* uses correlated variables sampled from Normal distributions based on the SNP correlation structure of chromosome22. Chromosome 22 estimates are from GENEVA-ARIC European-American data as in ref. [6]. See the *readme.pdf* file in the **GD** repository for details on the scripts used to generate the mock data.

References

[1] Donoho, D.L., Tanner, J.: Sparse nonnegative solution of underdetermined linear equations by linear programming. Proc Natl Acad Sci USA **102**(27), 9446–9451 (2005)

[2] Donoho, D.L.: High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. Discrete Comput Geom **35**(4), 617–652 (2006)

[3] Donoho, D., Tanner, J.: Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. Philos Trans A Math Phys Eng Sci **367**(1906), 4273–93 (2009)

[4] Donoho, D.L., Maleki, A., Montanari, A.: The noise-sensitivity phase transition in compressed sensing. IEEE Trans Inform Theory **57**, 6920–6941 (2011)

[5] Candès, E.J., Plan, Y.: A probabilistic and RIPless theory of compressed sensing. IEEE Trans Inform Theory **57**(11), 7235–7254 (2011)

[6] Vattikuti, S., Lee, J., Chang, C., Hsu, S., CC, C.: Application of compressed sensing to genome wide association studies and genomic selection. GigaScience (in review)

[7] Donoho, D.L., Maleki, A., Montanari, A.: Message-passing algorithms for compressed sensing. Proc Natl Acad Sci USA **106**(45), 18914–18919 (2009)

[8] Donoho, D.L., Tanner, J.: Precise undersampling theorems. Proc IEEE **98**(6), 913–924 (2010)

[9] Vila, J., Schniter, P.: Expectation-maximization gaussian-mixture approximate message passing. IEEE Trans. Signal Process **61**, 4858–4672 (2013)

[10] Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., Visscher, P.M.: Common SNPs explain a large proportion of the heritability for human height. Nat Genet **42**(7), 565–569 (2010)

[11] Vattikuti, S., Guo, J., Chow, C.C.: Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet **8**(3) (2012)

[12] Vattikuti, S., Chow, C.C.: Software: Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. https://github.com/ShashaankV/MVLME/

[13] Lee, J.J., Chow, C.C.: Conditions for the validity of snp-based heritability estimation. Human Genetics (2014)

[14] Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann Appl Stat **1**(2), 302–332 (2007)

[15] Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J Stat Softw **33**(1), 1–22 (2010)