A PRIMER ON GENOTYPE IMPUTATION

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Professional Studies in Agriculture and Life Sciences

Field of Integrative Plant Science, Specializing in Plant Biotechnology

by

Xavier Lopez

May 2021

ABSTRACT

Haplotype phasing and genotype imputation have become commonly used for genetic studies of all types, but especially for genome wide association studies and as a tool in plant breeding programs. As the ability of labs of all sizes to implement genome wide association studies increases, the skills and knowledge needed for preparing these data must be more commonly taught to students through all levels of higher education. This document is meant to serve as an introduction to the underlying concepts and show an example of the basic implementation and troubleshooting of haplotype phasing and genotype imputation for genetic studies.

BIOGRAPHICAL SKETCH

Xavier Lopez was born in New York City and grew up in Brooklyn, NY. In 2015, Xavier earned a Bachelor's Degree in Biology with a concentration in Neuroscience, as well as a Certificate in Biotechnology from University of Rochester. After graduating, Xavier worked for several years at the American Museum of Natural History teaching the public about the biology and ecology of Lepidoptera as well as helping care for the various animals that call the museum home. While at the AMNH, Xavier audited classes at the Richard Gilder Graduate School to learn various programming languages and started to develop an academic interest in Plant Science through non-fiction books about the natural and cultural histories of crops.

Xavier moved to Ithaca, NY with his partner in 2017 when she began her Doctor of Veterinary Medicine degree at Cornell University. While in Ithaca he worked as an Instrument Specialist at Q^2 Solutions, where he maintained, repaired, and validated equipment for a LCMS laboratory that specialized in clinical trial testing. During his time at Q^2 Solutions, he audited several plant science courses and then applied to the Master of Professional Studies program at Cornell University. In 2020, Xavier began working towards a degree in Integrative Plant Science specialized in Plant Biotechnology under the guidance of Professor Michael Scanlon.

ACKNOWLEDGMENTS

I would like to thank Savanah Marie Dale and Dr. Michael Gore for letting me take part and contribute to this project.

I give thanks to my mother for teaching me it's never too late to go back to school. Thanks to my father for teaching me to always "read the freakin' manual" and giving me the confidence to engage with problems I've never encountered before. To my partner (and soon to be wife) - thank you for your unwavering support both before and during my time here at Cornell.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

.vcf: Variant Call Format

.vcf.gz: Variant Call Format (GNU-zipped)

DR^2: Estimated squared correlation between estimated allele dose and true allele dose

EM: Expectation-maximization

GWAS: Genome Wide Association Study

HMM: Hidden Markov Model

LD: Linkage Disequilibrium

NAM: Nested Association Map (or 'Mapping')

QTL: Quantitative Trait Locus (or 'Loci')

SNP: Single nucleotide polymorphism

eQTL: expression Quantitative Trait Locus (or 'Loci')

Introduction to Imputation: Usage & History

Imputation Basics: What Is It, Why Use It?

In the vast majority of situations, access to knowledge of what nucleotides are present at most loci of a genome is simply not feasible. The creation of a reference genome is in itself a significant endeavor that requires resources unavailable to most labs, and any given reference genome will not necessarily capture the full extent of variation within a species. The time and cost of genotyping has fallen precipitously in the past 20 years, but the level of read coverage needed for full genome assemblies is still not viable for projects like plant breeding programs or GWAS, where the number of individuals to be genotyped is often in the hundreds or thousands.

To get around these limitations, haplotype estimation and genotype imputation are now standard practice in many genetics studies. At their core, these techniques use a higher read coverage reference panel to infer missing information from lower read coverage sequences in another individual or set of individuals (Marchini & Howie, 2010). More specifically, the reference panel is composed of individual haplotypes that have been sequenced at many SNPs with high read coverage (relative to the study individuals). Through the haplotype estimation step, also known as phasing, study individuals' SNPs are used to estimate which segments of their genomes correspond to haplotypes in the reference panel (Marchini & Howie, 2010). Once the haplotypes of the study individuals have been estimated, a statistical model can be applied that will try to find the most likely allele at any SNP that was in the reference panel, but not in the study individual's original genotyped sequence (the imputation step). It is

important to recall that these are estimates and therefore come with a certain amount of uncertainty; however it is well worth the uncertainty considering the astronomical costs of achieving a similar level of coverage otherwise.

As costs for sequencing have fallen, the use of genotype imputation has allowed for a simultaneous expansion in the number of individuals that can be genotyped for a project and the level of marker coverage that each individual can be sequenced to, which in turn significantly increases the power of methods like GWAS that benefit from larger sample sizes. Additionally, the statistical methods used in imputation pipelines allow for the use of datasets that were originally genotyped using different DNA microarrays, which aids with setting up reference panels as well as allowing for more data transferability between experiments (Zhou et al., 2017). This transferability also allows for use of population-specific reference panels to help identify rare or low-frequency variants (Mitt et al., 2017). Imputation is not just for finding missing genotype data in the context of GWAS though - the same methods can be applied to increase the precision of QTL-mapping, to help find non-SNP variation (such as copy number variants), or as an additional step to reduce genotyping methods error rates (Marchini & Howie, 2010).

History of Imputation in Genetics

Like many concepts in quantitative genetics, the theory behind genotype imputation has existed for significantly longer than the technical ability to successfully do so. The concept of maximum-likelihood estimators originated in the late 17th century but was first popularized in genetics by Ronald Fisher in the 1910's and in his seminal work *Statistical Methods for Research Workers* (Fisher, 1925). By the 1970's,

statisticians were starting to develop algorithms for maximum-likelihood methods, which would later form the basis of haplotype estimation and genotype imputation (Dempster et al., 1977). The first among these was the EM algorithm, an iterative method to find the most likely parameters in a statistical model where the equations cannot be solved directly due to missing data. While the EM algorithm and others like it provide good estimators, at that time the larger issues in using imputation for genetic sequence prediction were the sparseness of actual genetic data (due to very high sequencing costs) and the computational power needed to derive estimators using iterative methods.

With the development of high-throughput sequencing in the 1990's and rapidly decreasing computation costs, genetic markers became more commonly used in studies across the fields of biology, although most labs still could not afford to sequence at their desired read coverage. This in part led to the formation of the HapMap project, which in 2005 published the first human haplotype maps for general use to the scientific community (Altshuler & Donnelly, 2005). The first haplotype map for maize was soon after, which involved the genotyping of parent lines used in the creation of a nested association mapping population (Gore et al., 2009). The most recent version was used as the reference panel for this project (Bukowski et al., 2018). With the development of haplotype maps such as these for use as reference panels, the usage of GWAS for both trait mapping and plant breeding programs has risen dramatically over the past decade.

An Example of Imputation for Speed Breeding

Outline of the Project

The goal of this project is to combine imputation with speed breeding to substantially increase the rate of genetic gain in a sweetcorn breeding program. So far, a sweetcorn association panel of several hundred lines has been composed to compare against a combination of larger diversity panels, which was used to perform a GWAS seeking causal variants for nutritional content. These GWAS results were then used to create a model that chose 9 parental lines that form the program's F_2. Because there are so few parental lines, they can be sequenced to a very high density, which will allow for the building of a Practical Haplotype Graph that can be used for imputation during the speed breeding portion of the program (Buckler Lab, n.d.).

The F2 onwards will be grown using speed breeding, which is a set of techniques meant to reduce the intergenerational time of the program significantly. This is done through a combination of a controlled environment with artificially lengthened daylight, early seed harvesting, and artificial seed drying, among other possible factors, and allows for production of 4-6 generations per year instead of 1-2 (Watson et al.,2018). These F_2 populations will be narrowed down to 5 $F_{2:3}$ populations, which will then undergo single seed descent for another few generations. The derived lines' genotypes will be imputed using this same pipeline but adapted to work with the practical haplotype graphs, which will impute genotypes for use in a genomic selection model, bypassing the need to grow plants to maturity for phenotyping. In short, this should allow for a program to progress from selecting F_2

parents to planting F_5 field trials in one to two years. The component of the project outlined below is an elaboration of the earlier steps, specifically that this script will be adapted into the Practical Haplotype Graph for the speed breeding component of the project.

How the Imputation Was Done

The imputation was done on a dedicated lab server running a standard environment used by the Cornell Computational Biology Service Unit. Third party software packages were used: BCFtools version 1.11, Tabix version 1.11, Beagle version 5.0, and vcftools version 0.1.16 (Danecek et al., 2021; Browning et al., 2018; Danecek et al., 2011). Imputation of SNPs in the study data was performed as follows:

1) Downloading and preprocessing of data and reference panels
2) Formatting of the haplotype map components
3) Formatting of the study individuals data
4) Generation of the completed reference panel
5) Imputation

For a student who has never done genotype imputation before, this might seem like a surprising amount of formatting and preprocessing before the actual imputation step (which using modern programs, is actually relatively simple to code). When working with genotype data, especially at scales seen in modern breeding programs, formatting is paramount to be able to align or compare sequences. Much of this process involves consideration of the experimental design and origin of the data, as can be seen in figure 1, where most steps involve the filtering or reformatting of the original data.

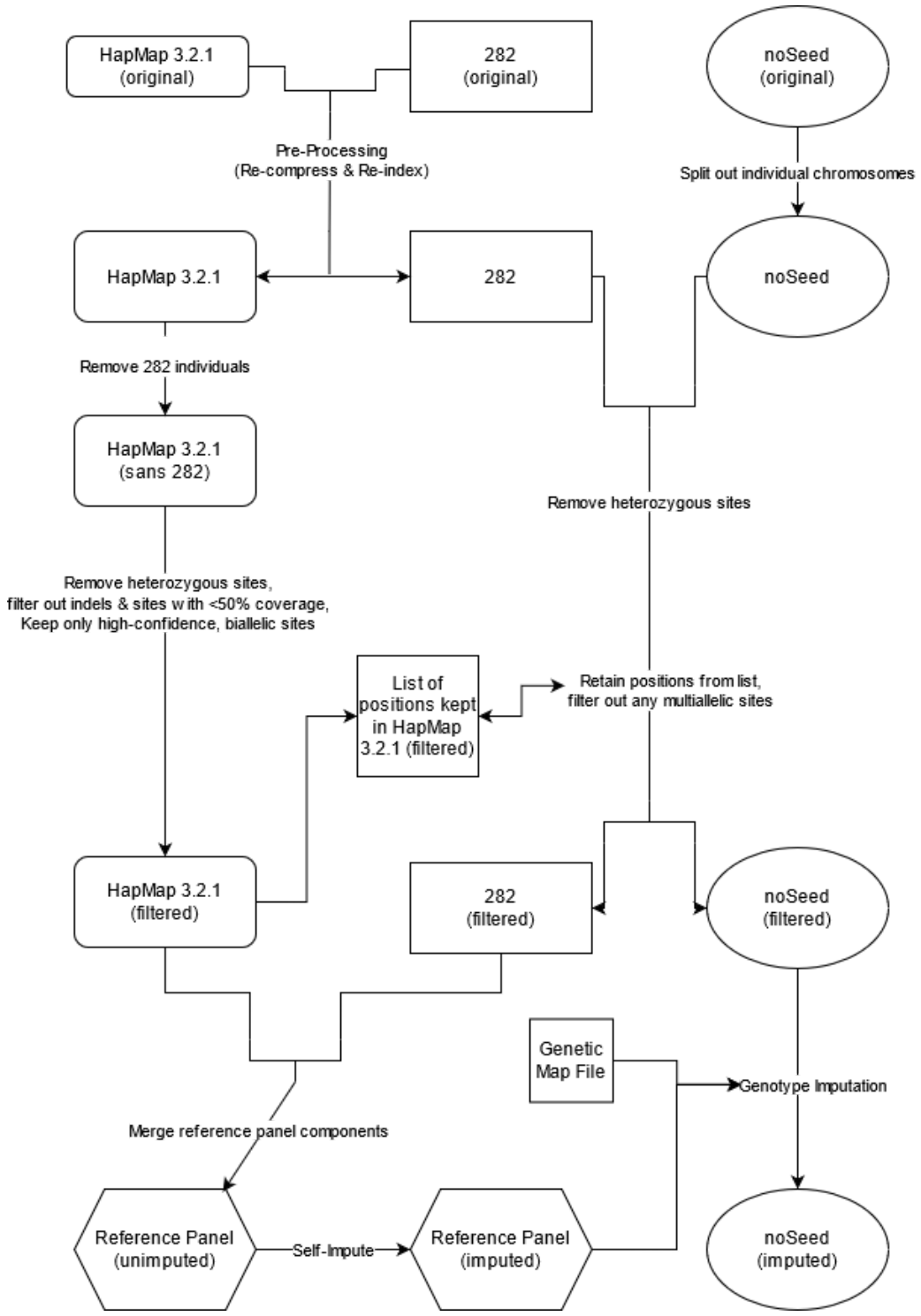

Figure 1. This flowchart shows the steps involved in preparing the data and imputing the SNPs of the noSeed_denovo_sweetcorn dataset. Files are shown as nodes, while filtering, estimating, and imputation steps are annotated on edges.

Downloading and preprocessing of data are relatively simple steps that will vary from project to project. In this case, the data to be retrieved were used to generate the reference panel in step 4 - the maize HapMap 3.2.1 (unimputed), and the maize 282 association panel found on the Panzea project Cyverse server. HapMap 3.2.1 is composed of whole genome sequences from 1,210 maize lines and has approximately 83 million SNPs, while the 282 panel has the same markers from a subset of lines that had higher coverage sequencing data (Bukowski et al., 2018). The preprocessing for these two files screen for errors in downloading, such as a truncated or corrupted copy of the file and re-compresses and re-indexes the .vcf.gz files. The preprocessing of the study genotype data is similar, but must also separate genomes into chromosome specific files. This isn't necessary per se (so long as the reference panel files are formatted the same), but having each chromosome as a separate file allows for faster compute times and easier troubleshooting if there are errors downstream.

The formatting of the haplotype map will vary by project, but generally the goal of this step is to make sure the haplotype map is formatted the same as the study data. The haplotype map will be more thorough than the study data, so this step essentially filters out unwanted information. In this instance, the 282 panel is also a subset of the HapMap 3.2.1 data, so it is important to remove any individuals in the former from the latter to prevent double-counting. Once there is only one record per individual, the HapMap 3.2.1 data is filtered to remove indels and only include high-confidence SNPs (approximately 30M of the 83M total). These SNPs would only be kept if they were homozygous, missing in fewer than half of the genotyped individuals, were biallelic, and had a minor allele frequency above 1%. These

parameters can change from project to project, so make sure to consider what the intention for the dataset is when formatting the haplotype map. The majority of other header information in the file (read depths, genotype posterior probabilities, etc.) is irrelevant to later imputation steps and are removed at this stage.

Once the HapMap 3.2.1 data is formatted, a file can be made that contains just the chromosome number and position of each SNP that has been retained. To format the 282 panel, once the unnecessary header information and heterozygous sites have been removed, it is filtered to keep the loci listed in the position file generated from the formatted HapMap 3.2.1 data. The study data is similarly formatted, but again filtering to only keep biallelic sites (it is not a safe assumption that a site that was biallelic in the reference panel will be biallelic in your data set).

The merging of the reference panel components combines the HapMap 3.2.1 and 282 datasets. The script also includes optional commands to ensure data is not altered in the merge and that the general alignment of the sequences is maintained. At this point the reference panel has a full list of all the biallelic sites that have met the previous criteria, but still has gaps in the sequence.

The imputation step is composed of two parts, both of which use Beagle. Beagle uses an imputation method based on a Hidden Markov Model (HMM). This model, given a set of reference alleles and haplotypes, tries to find the most likely haplotype for an individual by iteratively estimating and resampling haplotypes while taking the most parsimonious solution at each step. Because it is iterative and only accepts “better” solutions, it eventually converges on a most likely set of solution parameters, but it is generally very time and computationally intensive. Even on a

relatively powerful computer this process can take hours or even days depending on the dataset. The first part of imputation here is to perform the haplotype estimation (phasing) in the reference panel. Because they are from a NAM population and thoroughly genotyped, the 282 and HapMap 3.2.1 data are compared to estimate the haplotypes that make up the reference panel, producing the haplotype map. The second part of the step uses that haplotype map of the reference panel to impute the SNPs in the study data. This second imputation step uses this output and a genetic map file derived from the NAM population, which contains information anchoring genetic distances to physical distances on the chromosome. Beagle can then use this data to determine what the most likely SNP is in a given stretch of sequence for the study data, given that that stretch is of a certain haplotype. Once complete, the final step generates an index for the imputed file and extracts the positions, major allele frequencies, and estimated squared correlations between the estimated allele dose and the true allele dose (DR^2).

This final dataset will undergo a few quality assurance steps once compiled, such as filtering to only keep loci with a major allele frequency $\geq 1\%$ and $DR^2 \geq 0.8$, as this is ultimately going to be used for a breeding program. These data will be used as the basis for at least two different models. The first is a mixed linear model GWAS using all the SNPs that have been imputed and retained so far. The second is a multi-locus mixed model GWAS, which requires kinship coefficients (estimated using markers that are not in LD with each other) and filtering out of markers that are in full LD ($r^2 > 0.99$). This can give a better idea of causative SNPs for highly additive or polygenic traits, which are important for plant breeding decisions.

Tips for Performing Your Own Genotype Imputation

Knowing Your Tools

In the above example, multiple third party software packages were used to manipulate the data. It is important to become familiar with the software packages' capabilities before starting on the scripting, so as to know what each is capable (or incapable) of. The best resource is the software package's website, particularly any How-To's or Vignettes that are available on it. Seeing how a piece of code works in action will always help with trying to figure out how to tackle a problem, and many of these sites will also provide links to papers where the software was used for various manipulations or analysis, which can be helpful resources when trying to implement one's own code. This will also give the opportunity to see if maybe there are better software packages available for what you are specifically looking to do. While software like Tabix is relatively simple and used for basic manipulation of files, the software that actually performs filtering, modeling, or analysis tends to leave a lot of choice up to the user. At the moment, there's roughly a dozen different packages for haplotype estimation and genotype imputation, each with a different set of functions and utilities, as well as different methods of performing them, which can have an impact on your final dataset (Moorthy et al., 2019).

Additionally, it is always advisable to use actively supported versions of the software (although note that there are instances where the data being worked with may require you to use older versions). That said, just because something is deprecated does not mean it will not work, simply that one must be mindful of its limitations and bugs. For example, in this project, vcftools was used for several filtering steps, even

though it has been officially replaced by BCFtools. This led to many warning messages early in the project, though through digging into old forum posts, we were able to elucidate that it was actually just a formatting quirk in the headers that vcftools perceived as an error, and it could be safely ignored. The fact that vcftools was deprecated software meant this was a bit harder and had never been fixed since 2015, which leads to the final piece of advice.

Troubleshooting Advice

When in doubt, read the manual. When presented with an unusual error message, search engines are almost always able to find someone who has had the same problem in the past. If the problem is specific to a piece of software, check its website, StackOverflow (or similar software forums), or its Github repository (if it has one), for information from other users on how they've dealt with the issue.

If the problem is not one that produces a clear error message, oftentimes the best way to assess your script is to make a stripped down version that you can run step by step. For example, just using the smallest chromosome and purposely setting the burn in/iterations to 1 for imputation or phasing steps can let you test a script in just a few minutes as opposed to hours if you were to run it normally. Most of the software involved generates log files, which can be immensely helpful in identifying issues, but when it doesn't, it can be just as viable to make your own using the grep command. If a problem seems "untraceable" to a specific step, consider starting from the top of the script and looking at your input files - "garbage in, garbage out" as they say. And finally, if all else fails, go for a walk: sometimes (especially with coding) a solution is easier to find with a fresh pair of eyes and some time away from the keyboard.

WORKS CITED

Altshuler, D., Donnelly, P., The International HapMap Consortium (October 2005). "A haplotype map of the human genome". *Nature*. 437 (7063): 1299–1320. doi:10.1038/nature04226. ISSN 1476-4687.

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338–348. doi:10.1016/j.ajhg.2018.07.015

Buckler Lab (n.d.). Practical Haplotype Graph. Retrieved April 29, 2021, from https://www.maizegenetics.net/phg

Bukowski, R., Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, Dawen Xu, Bicheng Yang, Chuanxiao Xie, Longjiang Fan, Shibin Gao, Xun Xu, Gengyun Zhang, Yingrui Li, Yinping Jiao, Doebley, J. F., Ross-Ibarra, J., Lorant, A., & Buffalo, V. (2018). Construction of the third-generation Zea mays haplotype map. *GigaScience*, *7*(4), 1–N.PAG. doi:10.1093/gigascience/gix134

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi:10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). doi:10.1093/gigascience/giab008

Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society, Series B*. 39 (1): 1–38. JSTOR 2984875. MR 0501537

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd (Edinburgh). ISBN 978-0-05-002170-5.

Gore, M. A., Chia, J.-M., Elshire, R. J., Qi Sun, Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H., & Buckler, E. S. (2009). A First-Generation Haplotype Map of Maize. *Science*, *326*(5956), 1115–1117. doi:10.1126/science.1177837

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 499–511. doi:10.1038/nrg2796

Mitt, M., Kals, M., Parn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Magi, R., & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *EUROPEAN JOURNAL OF HUMAN GENETICS*, *25*(7), 869–876. doi:10.1038/ejhg.2017.51

Moorthy, K., Jaber, A. N., Ismail, M. A., Ernawan, F., Mohamad, M. S., & Deris, S. (2019). Missing-Values Imputation Algorithms for Microarray Gene Expression Data. *Methods in Molecular Biology (Clifton, N.J.)*, *1986*, 255–266. doi:10.1007/978-1-4939-9442-7_12

Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M.-D., Asyraf Md Hatta, M., Hinchliffe, A., Steed, A., Reynolds, D., Adamski, N. M., Breakspear, A., Korolev, A., Rayner, T., Dixon, L. E., Riaz, A., Martin, W., Ryan, M., Edwards, D., … Hickey, L. T. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature Plants*, *4*(1), 23. doi:10.1038/s41477-017-0083-8

Zhou, W., Fritsche, L. G., Das, S., Zhang, H., Nielsen, J. B., Holmen, O. L., Chen, J., Lin, M., Elvestad, M. B., Hveem, K., Abecasis, G. R., Kang, H. M., & Willer, C. J. (2017). Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genetic Epidemiology*, *41*(8), 744. doi:10.1002/gepi.22067