Sequencing Millions of Animals for Genomic Selection 2.0

J.M. Hickey[1], **G. Gorjanc**[1], **M.A. Cleveland**[2], **A. Kranis**[1,3], **J. Jenko**[1], **G. Mészáros**[1], **J.A. Woolliams**[1], **and M. Perez-Enciso**[4]

[1]The Roslin Institute and R(D)SVS, The University of Edinburgh, UK, [2]Genus PLC, Hendersonville, USA, [3]Aviagen Limited, Edinburgh, UK [4]Centre for Research in Agrigenomics (CRAG) Bellaterra; and ICREA, Barcelona. Spain

ABSTRACT: Genomic selection has high economic value in breeding programs and this value will result in large data sets of genotyped and phenotyped individuals being generated. With appropriate developments in sequencing methods and strategies, and bioinformatics and imputation technologies these large data sets could be sequenced. That may enable larger proportions of the genetic variance to be finely mapped to causal variants. Finely mapping many causal variants will open up several new opportunities for breeding animals such as use of inflated recombination rates, genome editing, accurate estimation of breeding values in crosses, and capitalization of *de-novo* mutations. This paper uses simulation to illustrate how breeding programs may capitalize on these opportunities: hence genomic selection 2.0.

Keywords: genomic selection 2.0; genome editing ; recombination

Introduction

Genomic selection (**GS**) has absorbed a great proportion of research effort in animal breeding in recent years. Consequently, many breeding programs have now incorporated GS, improving accuracy in selection decisions about young individuals. This success, together with dwindling sequencing costs, justifies the collection of genomic information on huge numbers of individuals. Major breeding programs have the capacity to genotype more than one hundred thousand individuals per annum, which means that more than a million individuals with genomic information could be assembled within a decade, in many of such programs. To date the major benefit derived from genomic information has been more accurate predictions of estimated breeding values (**EBV**) for young selection candidates and these accuracies may now be approaching their upper limit, at least when well-designed and large training populations are available. To capitalize more fully on the scale of the data generating capacity that is now possible, new ways to select individuals are needed which go further than maximizing accuracy of EBV. The objectives of this paper are to: (i) review the extant GS paradigm; (ii) present a vision for how sequence information could be generated for huge livestock data sets (e.g. millions of individuals); and (iii) describe a few approaches through which animal breeders could utilize this capacity for increasing the rate of genetic progress.

Three Classifications for Genomic Selection

GS0.0 - When GS was proposed, its underlying assumption was that linkage disequilibrium information between causal variants and high-density markers spread across the whole genome would enable accurate predictions of EBV. This could be referred to as GS0.0. With training populations of not more than a few thousand individuals the model underpinning GS0.0 worked well for traits controlled by a small number of causal variants, each with a large effect.

GS1.0 - Almost all traits of agronomic importance are controlled by many causal variants in complex ways. Therefore the GS that has worked in practice has primarily used genomic information to capture large linkage blocks that are shared by closely related individuals. This could be referred to as GS1.0. In GS1.0 the effects of individual causal variants are neither accurately estimated nor finely mapped (this is not their objective, in any case). However, if sufficient numbers of very close relatives to the selection candidates are present in the training population, then accurate predictions of EBV can be obtained with a relatively small number of markers (e.g. <10,000). However, the accuracy of these predictions quickly dissipates as the relatedness between the training population and prediction candidates decreases.

GS2.0 - GS2.0 could be used to describe the type of GS that will be transitioned to in the next years. Because of recent and probable future technological advances in sequencing methods and imputation GS2.0 will likely be underpinned by the ability to have sequence information for all candidates and huge training populations (e.g. all breeding and production individuals). With such huge data sets, larger proportions of the total genetic variance will be ascribed to causal variants, leading to new opportunities to select animals, which go beyond more accurate and more persistent EBVs.

For example: (i) Genome Editing (**GE**), which was the Nature Method of the Year in 2011, could be used to create new variation or 'repair' old variation in targeted ways; (ii) breeding programs could be modified to utilize subset of the millions of naturally occurring de-novo mutations that might be agronomically valuable; or (iii) recombination rate could be increased (through selection, GE, or by manipulating the environment) and this could enable more of the standing genetic variation be utilizable in each generation. Additionally breeders may be able to capitalize on different kinds of variability, not only single nucleotide polymorphisms (**SNP**), but also copy number variants or other variants, use of functional information, prediction of merit in crosses, and less investment in phenotyping by getting more stable EBVs.

Desirable Properties of GS2.0 Data Sets.

GS1.0 training populations needed to comprise lots of close relatives of the individuals for whom EBV were to be predicted, so that the effects of large shared haplotypes could be estimated well and so that there would be a limited number of recombinations, to break apart these haplotypes, between the training and prediction individuals. GS2.0 training sets will need to be different to capitalize on increased volumes of genomic data. In summary, the training set will need to be huge and comprise of individuals that are as unrelated to each other as possible. Such data will be needed because: (i) there is a huge number of segregating variants that need to be estimated (e.g., tens of millions SNP, millions of Indels, etc); and (ii) because in small data sets these segregating variants are highly correlated due to recombination being a very rare occurrence. Huge data sets of unrelated individuals not only enable utilization of the recombinations that occur in each individual within the data sets, but also all of the ancestral recombinations in the large coalescent tree that conceptually connects all individuals. With a sufficient number of phenotypes and recombinations, fine mapping the causal variants for large proportions of the genetic variance would be more powerful than it is today.

Generating Huge Data Sets with Sequence Information

Genotype imputation, coupled with complementary genotyping strategies, has been central to the success of GS1.0, because these enabled lots of individuals to be genotyped with the required density at low cost. These methods and strategies capitalized on some of the features of livestock populations: (i) small number of sires; (ii) large family sizes; (iii) abundant pedigree information; and (iv) high degree of relatedness between individuals. In reality, cheap and accurate genotype information on target individuals (e.g. candidates and training individuals) was obtained for GS1.0 by genotyping the male ancestors at high density, phasing their whole chromosomes, and then using low-density markers to detect the recombination locations and track the inheritance of these haplotypes through any intermediate female ancestors. Therefore, by way of example, in a pig breeding program 50,000 markers can be imputed in selection candidates with an accuracy of >0.96 (i.e., correlation between true and imputed genotype) at a cost of <\$20.38 (assuming the cost of genotyping with 50,000 markers is <\$80, and with 384 markers is <\$20 (Huang et al. (2012)).

This approach will be suboptimal for GS2.0, primarily because the cost of generating high-quality sequence information on all male ancestors is currently prohibitively high, to enable their offspring to be genotyped with standard low-density platforms. For example, if we assume that the library preparation step costs \$100, and 1*x* of sequence for one individual costs \$100, then the total cost of sequencing the 500 sires that might be used per annum in an elite pig nucleus at 30 *x* would be \$1.5 million. In comparison, genotyping these 500 sires with 50,000 SNP would currently cost less than \$40,000, if we assume a chip with 50,000 SNP costs <\$80.

An alternative approach could be to sequence all individuals at low-coverage, use an imputation-like approach to build consensus haplotypes to finally impute the high-coverage sequence. (NOTE: Imputation algorithms, together with suitable sequencing strategies can even detect de-novo mutations, albeit with a time lag of two or three generations). With this strategy, the actual haplotypes of individuals would be sequenced at low-coverage using the individuals' own DNA, but high-coverage sequencing of these haplotypes would be obtained by accumulating all of the low-coverage reads from all the individuals that carry the same haplotype. This approach has three potential advantages over the approach that uses key ancestors sequenced at high-coverage. Firstly, it seems inefficient to sequence the same haplotype at high coverage multiple times in different individuals. Secondly, accurate imputation involves the resolving of the combination of haplotypes that an individual carries and in the era of sequence data the refinement of recombination locations will be the most important component of this. For an individual to be imputed, low-coverage sequence data can generate many more markers than SNP chips for the same cost. This vastly greater number of markers will allow the recombination locations to be much more precisely located. Thirdly, in GS1.0 accurate genotype imputation was not possible unless very close ancestors (e.g. sires and grandsires) of individuals were genotyped at high-density. Sequencing only the key ancestors implies that not all sires and grandsires will be sequenced at high-coverage, due to financial constraints, and this would severely weaken the power to do imputation downstream. In Box 1 simulation is used to evaluate the power of these two different approaches and the results confirm the advantage of using low coverage sequencing on many individuals.

Sequencing Sperm Cells to Increase Imputation Accuracy

The accuracy of imputation is partly affected by the accuracy with which the whole chromosomes of individuals are phased. In the absence of the ability to perfectly phase sequence data, it may be more powerful to capitalize on the high fecundity of sires in animal breeding programs and the haploid nature of sperm cells. Box 2 describes a simulation that was used to compare the power of sequencing a number of sperm cells of a sire versus sequencing the sire itself. Sequencing sperm cells results in highly accurate phasing of the genome of sires, which in turn leads to highly accurate imputation of the sire gametes into the progeny. However, when focusing all of the sequencing effort on sperm cells, the low levels of recombination during meiosis result in part of the gametes of the sire not being represented in sequenced sperm cells. This leads to inaccurate imputation for these regions in comparison to the accuracy of imputation when sequencing the sires DNA directly. Hence, an optimal approach might be to utilize sequencing resources on selection candidates and once males are selected their sperm cells could be sequenced to boost the accuracy of their phasing. Sires would therefore never be directly sequenced with high coverage.

Box 1. Evaluation of the power of low-coverage sequence strategies in comparison to the "key ancestors approach" to enable accurate imputation.

Simulation: One chromosome 1 Morgan in length, 25 sires, 1000 progeny to be imputed.

5 scenarios

1. Sequence sires at 40*x* (1000*x* in total) and genotype progeny with 20 low-density SNP markers
2. Sequence sires at 40*x* (1000*x* in total) and genotype progeny with 200 low-density SNP markers
3. Whole genome[1] SNP on sires and genotype progeny with 20 low-density SNP markers
4. Whole genome[1] SNP on sires and genotype progeny with 200 low-density SNP markers
5. Do not sequence sires but sequence 1000 progeny at 1*x* (1000*x* in total)

[1]Whole genome SNP means that there are no errors or uncertainty in SNP calls due to the sequencing depth

Results

Table 1. Correlation between true and imputed sequence

Scenario	Correlation
1	0.678
2	0.678
3	0.752
4	0.887
5	0.921

Conclusions

Imputing sequence (even at 40x) is harder than imputing SNP (Sc. 1 and 2 versus Sc. 3 and 4).

Spreading sequence across progeny is better than sequencing sires at high *x* (Sc. 5 versus Sc. 1 and 2) because marker density in progeny is higher.

Required Developments to Imputation Algorithms

Imputation algorithms that were developed for application in livestock were largely designed to use SNP genotype data that has high certainty in the genotype calls and a large proportion of data is missing in a structured way (e.g. Druet and Georges (2010); Hickey et al. (2011)). Low-coverage sequence data have uncertain or probabilistic genotype calls and the data is missing at random. Thus imputation algorithms will need to be completely re-written to utilize both heuristic and probabilistic methods; the first to ensure scalability for huge data sets and the latter to handle uncertain data. Recently, we have taken the first steps in developing such an algorithm (Crossa et al. (2013)). Focusing only on the heuristic component, our prototype algorithm was designed to impute the missing data from low-coverage genotyping-by-sequencing (**GBS**) in inbred individuals (NOTE: inbred individuals are phased *de facto*). The algorithm begins by identifying individuals that share a haplotype at a region on the basis that those that share haplotypes do not have any opposing homozygote loci within that region. After identifying these individuals their low-coverage reads are stacked and a high-coverage consensus haplotype is formed. This haplotype is then back imputed into all individuals that carry it. In a test data set of maize inbred lines each individual had on average 44% of the markers missing. After running the algorithm this reduced to 20%. Much of the remaining missing markers were missing due to mutations in the GBS restriction enzyme sites and thus were actually informative markers in themselves and thus not the target for imputation.

Box 2. The power of imputation when sequencing sperm cells.

Simulation: One chromosome 1 Morgan in length, 50 sires, 500 progeny to be imputed, 1500 sperm cells available

5 scenarios

1. Sequence sires at 30*x* (1500*x* in total) and genotype progeny with 20 low-density SNP markers
2. Sequence sires at 30*x* (1500*x* in total) and genotype progeny with 200 low-density SNP markers
3. Do not sequence sires; instead sequence 1500 sperm cells (30 per sire) at 1*x* (1500*x* in total) and genotype progeny with 20 low-density SNP markers
4. Do not sequence sires; instead sequence 1500 sperm cells (30 per sire) at 1*x* (1500*x* in total) and genotype progeny with 200 low-density SNP markers
5. Sequence sires at 10*x* and 1000 sperm cells (20 per sire) at 1*x* (1500*x* in total) and genotype progeny with 20 low-density SNP markers
6. Sequence sires at 10*x* and 1000 sperm cells (20 per sire) at 1*x* (1500*x* in total) and genotype progeny with 200 low-density SNP markers

Results

Table 2. Correlation between true and imputed sequence

Scenario	Correlation
1	0.704
2	0.756
3	0.761
4	0.933
5	0.783
6	0.967

Conclusions

Sequencing sperm cells is more powerful than sequencing sires.

Sequencing a mixture of sperms cells and the sires is the most powerful.

Power of Raw Low-Coverage Sequence Data

Currently there is no powerful imputation algorithm that has been explicitly designed for imputing low-coverage sequence data in livestock. Algorithms that have been designed for human data sets could be used in livestock. However, for classical imputation in livestock these algorithms have been shown to be very suboptimal, when measured using the correct metrics (i.e. the correlation). In the absence of powerful imputation algorithms low-coverage sequence data is still very powerful for genomic selection in livestock. The results of a simulation described in Box 3 show that when marker density is high (e.g., 300,000 markers) the accuracy of prediction with very low sequencing depth (e.g., 1*x*) can be as high as with high-coverage sequence or high quality SNP data. However, generating 300,000 markers with GBS technology at 1*x* coverage can be significantly cheaper than genotyping with 300,000 SNP markers.

Box 3. The power of raw low-coverage sequence data

Simulation

Thirty chromosomes each 1 Morgan in length, 1000 individuals in each of two generations, 50 males and 500 females from generation 1 chosen to be the parents of generation 2. Use 1000 individuals in generation 1 to predict the genomic breeding of individuals in generation 2 based on ridge regression. GBS or SNP markers with a density of 3000, 10000, 60000, or 300000. GBS had a range of sequencing depths from 0.01*x* to 20*x*.

Results

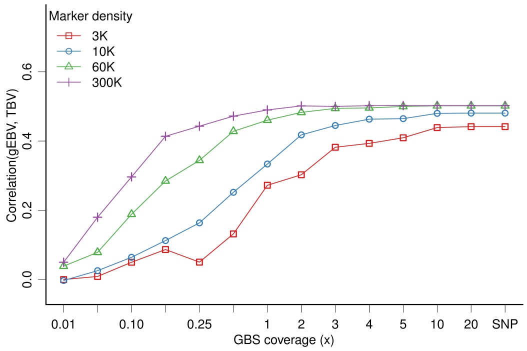

Figure 1. Accuracy of genomic selection when using SNP or a range of sequencing depths for GBS data

Conclusions

Low-coverage sequencing in the form of GBS is as powerful as SNP. When marker density is high (e.g. 300000 markers) coverage can be as low as 1*x* in the absence of imputation.

Required Developments to Sequencing Technologies

Despite the sheer amount of data delivered by current sequencing technologies, these will be by no means the end. Throughput of next generation sequencing (**NGS**) will continue to increase in the near future although a true change in paradigm should come from reliably sequencing long single DNA strands. This technology may have a dramatic impact to animal breeding, because it would allow directly observing phase and therefore increasing imputation accuracy, together with better characterization of structural variants. Within the parameter space of available sequencing technologies, some new techniques could be optimized, so that they complement the possibilities of imputation technologies more optimally. For example, the use of probes, restriction enzymes, multiplexing, and inference (e.g., DNA Sudoku (Erlich et al. (2009))) could be further optimized, so that the user has control to the degree to which a pair of individuals overlap in their resulting low-coverage sequence reads.

Discussion on the Steps in a Sequence Pipeline

It is important to remember that NGS data is not simply an increased number in SNP density as compared to SNP arrays. This is because SNPs on arrays are 'ascertained', i.e., they are discovered in a sample of individuals and chosen based on frequency (usually high MAF SNPs are chosen) and later these SNPs are genotyped in the sample under study. The effect of ascertainment in GS has not been studied in detail, although simulations show that SNP density and the demographic history of the species (primarily admixture events) have an important influence on how different will be genomic relationships computed with either ascertained SNPs or sequence (Perez-Enciso (2014)). A second, normally overlooked, reason that makes NGS data different from array genotypes is that the final set of NGS-based SNPs is very sensitive to the used options in the bioinformatics pipelines. For instance, standard SNP calling algorithms like samtools and GATK are biased towards the reference allele, especially at low depth. This may result in biased genotype calls. To remedy this, reference independent SNP calling algorithms are recommended (Nevado et al. (2014)). However, it is expected that the accuracy of genotype calling algorithms will increase as data sets increase in size.

Infrastructure and Human Resources Requirements

While sequencing costs are declining, this may not be so for the required bioinformatics infrastructures. First, NGS analyses are not, as of today, standard and completely automatized. This is resulting in a strong lack of specialized personnel with sufficient skills in both bioinformatics and genomic selection. Future animal breeding educational programs should take these two angles in consideration. Simultaneously, it is unlikely that raw NGS data generated throughout the world would be easily transferred with enough speed between labs. It is much more likely that bioinformatics analyses, once agreed and standardized, will be carried independently and then required information (e.g., variant files) merged. This process could be done iteratively with intermediate files being transferred via cloud services.

Deriving Breeding Benefit from Huge Volumes of Sequence Data

Genomic selection is now implemented in the major commercial breeding programs globally. The economic value of these programs will result in huge data sets being generated, and if advances in imputation and sequencing methodologies evolve sufficiently, it will be possible to generate sequence information for all of the individuals in such data sets. Analysis of such huge data sets will enable more of the genetic variation to be finely mapped directly and this could open up several new ways to increase the rate and sustainability of genetic gain in breeding programs including the use of de-novo mutations, the use of genome editing, and the use of inflated recombination rate.

Box 4. Use of genome editing for quantitative traits in livestock

Simulation

Ten chromosomes, each 1 Morgan in length. 50 generations of selection. Each generation consists of 500 males and 500 females. In each generation 50 sires and 500 dams were selected using genomic breeding values. Training set for genomic prediction consisted of the 1,000 individuals in the previous generation, genotyped with 20,000 SNP based on ridge regression.

Results

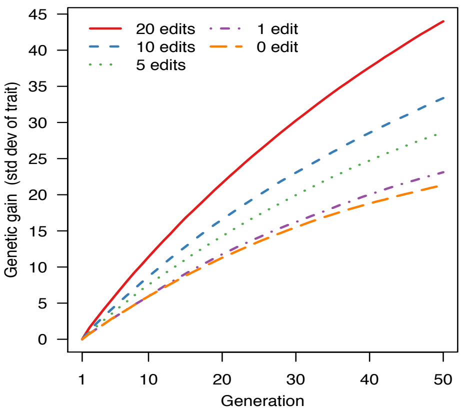

Figure 2. Response to selection when combining genomic selection to first select sires and then various numbers of genome edits on each of the selected sires.

Conclusions

Genome editing can complement genomic selection for quantitative traits. Further details are described in Jenko et al. (2014).

Using De-Novo Mutations in Breeding Programs

It is clear by now from selection experiments that exhausting genetic variability is almost impossible. Although several reasons have been invoked (epistasis among them), a plausibly important one is newly arising mutations. Every mammalian gamete is expected to bear ~30 new alleles (Kong et al. (2014)), i.e., the total amount of variation entering each year into any breeding scheme is non negligible. Most of these new variants will be lost, due to drift or selection against deleterious effects, but those that may have been targeted by selection will increase in frequency (in proportion to Ne × s, where Ne is effective population size and s, selection coefficient, i.e., the smaller the effective size the less effective selection is). These new variants will not be tagged by any SNP arrays; yet, they could be discovered with sequence data, in theory, and even imputed (see above). It can be hypothesized that, as sequence data is collected through generations, careful longitudinal analyses may reveal arising new mutants that could provide a competitive advantage to a specific breeding scheme.

Using Genome Editing in Breeding Programs

Genome editing (**GE**) is a technology that enables modification of genetic material in targeted ways (Niu et al. (2014); Cong et al. (2013)), for example single nucleotides can be targeted and modified with high accuracy. In breeding programs one use of GE could be to repair a small number of undesirable alleles in individuals that have otherwise high breeding values. Such an approach could make GE very complementary to GS. Individuals could be first selected on the basis of GS and then have some of their unfavorable alleles "repaired". The simulations in Box 4 show that using genome editing to "repair" even a modest number of unfavorable alleles in selected sires can give a major increase in the rate of response to selection in comparison to genomic selection, even for polygenic traits. However, to make GE work in practice for polygenic traits accurate fine mapping of causal variants is needed. Sequencing huge data sets may enable this.

Using Higher Recombination Rates in Breeding Programs

The amount of genetic variation available in a breeding program is affected by the number of causal variants, their frequency and size, and the degree to which they are linked to each other. The amount of genetic variation is one of the factors that affects the rate of genetic gain in a breeding program. During meiosis the rate of recombination is low and this limits the amount of standing genetic variation that is released for selection in each generation. Recombination is partially under genetic control (Kong et al. (2014)), thus it can be increased through selection. If carefully managed, this could lead to an increased response to selection. The simulations in Box 5 show that for a given selection intensity the rate at which genetic variation is exhausted is lower when the rate of recombination is higher. Using a grid search, a pair of scenarios was found (one with a high recombination rate and high selection intensity, the other with a lower recombination rate and lower selection intensity) in which the rate of loss of genetic variance across multiple generations of selection was almost identical. Because the scenario with a higher recombination rate could sustain higher selection intensity, the resulting response to selection across multiple generations was much higher. To make higher recombination rate work in practical breeding programs huge data sets are needed. Huge data sets are needed, because when recombination rate is higher, the genomic predictions can no longer depend on linkage and long-range correlations between markers and causal variants. With higher recombination rate, predictions would require that the effects of direct causal variants be much more finely mapped than is currently achieved.

Box 5. Use of increased recombination rate in livestock breeding

Simulation

One chromosome. 20 generations of selection. Genome length ranged from 1 Morgan to 10 Morgans, with selection intensity from 10% to 1.25%. A grid search was performed to find a pair of scenarios that had identical patterns of loss of genetic variation across the 20 generations of selection but different selection intensities and recombination rates.

Results

The scenarios with genome length of 1 Morgan with a selection intensity of 5% resulted in a loss of variance across 20 generations of selection that was almost identical to a scenario with a genome length of 10 Morgans and a selection intensity of 1.25% (Figure X). The higher Morgan scenario resulted in a much higher rate of genetic gain both in the short and long term.

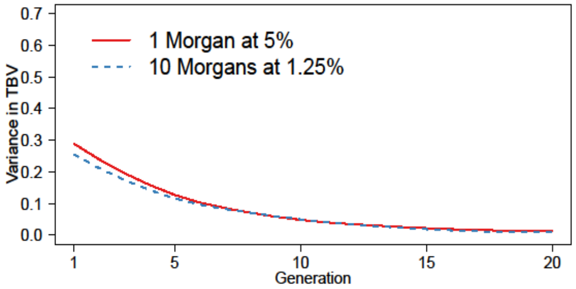

Figure 3. Loss of variance across generations

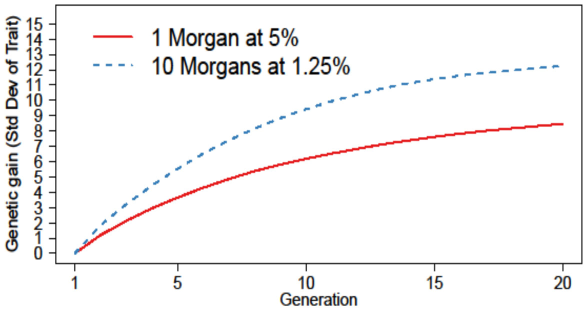

Figure 4. Response to selection across generations

Conclusions

Increasing recombination rate leads to faster genetic progress in the short and longer term. Further details are described in Mészáros et al. (2014).

Conclusions

We foresee that genomic selection 2.0 will dominate the future breeding industry. This implies availability of huge sequence datasets with the possibility of finely mapping large proportions of the genetic variation directly to its causal variants. To capitalize on this for faster and more sustainable genetic progress new breeding programs designs that make use of techniques such as increasing recombination rate, utilization of *de-novo* mutations, and genome editing will be needed. Genomic selection 2.0 will need new tools and techniques to cost effectively generate and analyze huge volumes of data. Single sperm sequencing, new ways of performing low-coverage sequencing and new imputation algorithms will need to emerge.

Literature Cited

Cong, L., F.A. Ran, D. Cox, S. et al. (2013). Science. 339:819–823.

Crossa, J., Beyenne, Y., Kassa, S. et al. (2013). G3. 6:1903-1926.

Druet, T., Georges, M. (2010). Genetics. 184:789-798.

Erlich, Y., Chang, K., Gordon, A. et al. (2009) Genome Res. 19:1243-53.

Huang, Y., Hickey, J.M., Cleveland, M.A. et al. (2012). Genet Sel Evol. 44:25.

Hickey, J.M., Kinghorn, B.P., Tier, B. et al. (2011). Genet Sel Evol. 43:12.

Jenko, J., Mészáros, G., Gorjanc, G. et al. (2014) Proc. Of the 10th World Congress of Genetics Applied to Livestock Production.

Kong, A., Thorrleifsson, G., Frigge M.L. et al. (2014) Nat. Genet. 46:11-16.

Nevado B, Ramos-Onsins, S., Pérez-Enciso M. (2014). Mol Ecol.

Niu, Y., B. Shen, Y. Cui, Y. et al. (2014). Cell. 156:836–843.

Mészáros, G., Gorjanc, G., Jenko, J. et al. (2014) Proc. Of the 10th World Congress of Genetics Applied to Livestock Production.

Pérez-Enciso, M. (2014) J Anim Breed Genet. doi:10.1111/jbg.12074