BIG IS BEAUTIFUL: BIOLOGY INFORMED SEQUENCE EXPLOITATION

M. Pérez-Enciso[1,2,3], M. Naval-Sánchez[1], J. Leno-Colorado[2] and A. Reverter[1]

[1] CSIRO Agriculture Flagship, Queensland Bioscience Precinct, St Lucia, QLD 4067, Australia
[2] Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain
[3] Institut Català de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

SUMMARY

We have entered the big data paradigm. Now that whole genome sequence data is available on a population scale basis, a fundamental issue is: what can be done with sequence data that cannot be achieved with former datasets? This question does not have a closed response, partly due to the fact that information contained in sequence data is highly repetitive (e.g., linkage disequilibrium) and also noisy (e.g., missing data due to shallow coverage). We argue that using accurate biology informed decisions can make a big difference in the prediction of genetic merit when sequence is available. Here we review the main kinds of external biological information and some approaches to combine these disparate sources. Despite the richness of resources available, two main difficulties lie ahead: (i) an improved understanding of the phenotype's biology to make the right the choice among the plethora of datasets available, and (ii) how this information is weighed and incorporated into selection decisions.

INTRODUCTION

The whole classical paradigm of animal breeding has been traditionally based on large datasets consisting of phenotypes and pedigree. Both kinds of information are rather homogenous and a unified, well accepted method was used for genetic evaluation, namely best linear unbiased prediction (BLUP). Molecular information in the form of low and high-density SNP arrays started to disrupt this data homogeneity. The amount of available molecular information in most livestock breeding programs has vastly increased recently, and this pace will only accelerate in the coming years. Today, the continuous decrease in sequencing and high performance computing (HPC) costs have made it conceivable the use of fully sequence in commercial breeding programs (Daetwyler *et al*. 2014).

Yet, it is important to realize that sequence data is not simply an increased SNP density. It is often said that, with sequencing, the causal mutations *are* in the data. But what is sometimes overlooked is that sequence data are very noisy, expensive to analyse, and error prone, especially at low coverage. As a result, derived genotypic data are highly unbalanced. For instance, in a large scale SNP discovery effort, where we analysed 120 pig genomes, only a few hundred SNPs out of all 45 million identified in total were called in all samples (Figure 1). This is to be the rule rather than the exception with this kind of data.

Despite initial enthusiasm based on simulation studies (Meuwissen & Goddard 2010), the limited empirical evidence on use of complete sequence for genomic selection so far calls for caution. Hayes *et al.* (2014) reported only a small (~4%) increase in accuracy compared to standard high-density array based selection. More recent simulations by Druet *et al.* (2014) and MacLeod, *et al.* (2014) suggest that the actual advantage will be heavily influenced by the allele distribution of causal variants and by recent demography (i.e. linkage disequilibrium). In parallel to the availability of larger genotype datasets and improved algorithms to predict genetic merit, vast amounts of new functional information are becoming available. After the sequencing of high quality reference genomes, gene expression datasets by RNA-seq across tissues are becoming available (e.g. Liao *et al*. 2014), and current and future essays on histone marks, methylation, open-chromatin transcription binding and chromatin conformation promise to unravel the regulatory landscape governing biological processes (Andersson *et al*. 2015). An advantage of this kind of information is that it can, partly, be transferred across species (Villar *et al*. 2015). For instance, metabolic pathways are well conserved across mammals or even across eukaryotes for fundamental pathways as well as gene expression levels (Brawand *et al*. 2011). On the contrary, the regulatory levels across mammals are highly dynamic (Schmidt *et al*. 2010; Villar *et al*. 2015). Here, we review the different sources of current and foreseeable available information, and we suggest that careful utilization of this biological information might boost genomic selection.

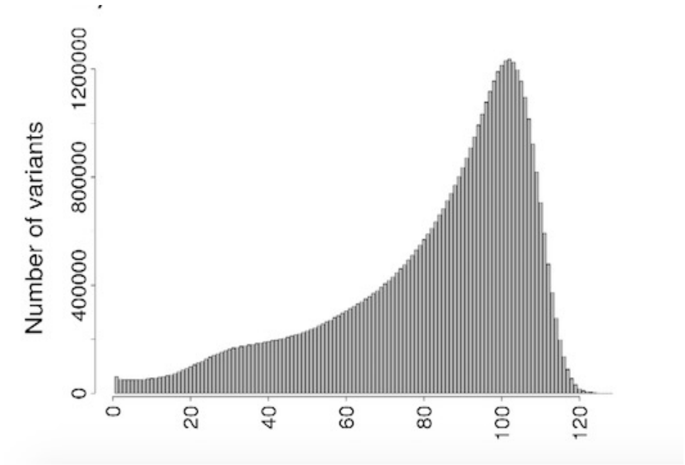

Figure 1: Number of individuals in which a given variant (SNPs) is observed. The data pertain to 128 pig genomes sequenced at varying depths, 4-20x, analysed with bwa and samtools. Figure from Bianco el al. (2015).

WHY FUNCTIONAL INFORMATION CAN BE USEFUL

Knowing the causal mutations is the holy grail for quantitative genetics. If these were known, much more accurate genetic predictions could be made, but note that this is but an extreme case of strong priors assigned to the SNPs available in the sequence dataset. In a recently published simulation study (Perez-Enciso *et al*. 2015), we showed that there is a clear law of diminishing returns when SNP density increases, and that the use of sequence would deliver only modest gains in accuracy. We predicted that only when using accurate biological information was sequence to pay off. Figure 2 shows our results. The two extremes are sequence data when used 'blindly', that is, without giving any different prior to any of the SNPs and inclusion of only the causal SNPs in the model. The latter strategy approaches an accuracy of 1, confirming our conjecture. Because all causal mutations are in the sequence, it is clear that wise choice of priors for each SNP can have a dramatic influence on prediction. Now, if all genes containing causal SNPs could be identified (red line) accuracy would increase by ~40%, as a result of disequilibrium with causal mutations. Yet, unfortunately, our simulations also show that miss- or incomplete specification of causal genes quickly diminishes accuracy (magenta and blue lines).

KINDS OF AVAILABLE INFORMATION

Table 1 shows a very shortlist of databases illustrating the wide diversity of data available that can be potentially used for improving the prediction of SNP functionality. These are: QTL, genome annotation, SIFT prediction, expression, methylation status, pathway information, gene ontology, among others. The new Functional Annotation of Animal Genomes (FAANG) consortium is currently gathering efforts to provide the same data to the animal genetics research community (www.faang.org) (Andersson *et al*. 2015) thus procuring a high quality genome annotation for domestic animals with unprecedented detail. Expression data are of particular relevance. Defining which genes are expressed in which cell types and developmental time points is fundamental to our understanding of development and disease. RNA-seq data pictures whole genome expression levels independently of the species/breed of interest. Coupling differentially expression analysis together with motif discovery or pathway analysis results in further insight into regulation and biology. Seminal studies comparing the regulatory landscape across vertebrates have proven that regulatory regions are highly dynamic with only a core being conserved across species (Schmidt *et al*. 2010; Villar *et al*. 2015). Therefore, the regulatory annotation of distinct tissues and developmental stages across domestic species is crucial for their study.

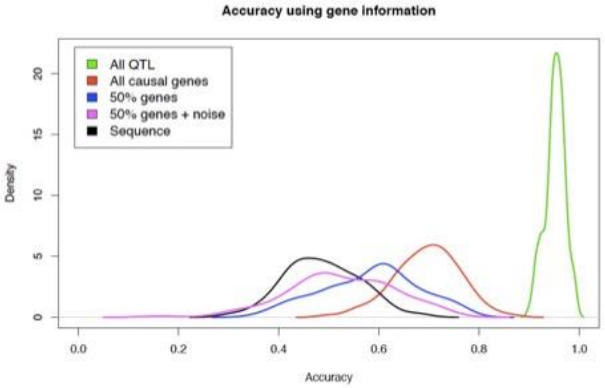

Figure 2. Comparison in accuracy between blind use of sequence (black) and if only QTL are employed in the model (green). The different curves show when either all SNPs within causal genes are used (red), 50% of the genes (blue) and added neutral SNPs (magenta). Details in Perez-Enciso et al 2015.

Related to this is the understanding of gene regulation itself. Motif discovery analysis on gene expression signatures is a popular alternative to detect regulatory regions and regulators in a given biological process. The underlying hypothesis to this strategy is that co-expressed genes tend to be co-regulated and therefore they might present similar Transcription Factor Binding Sites (TFBSs) in their regulatory region. Motif discovery analyses applied to differentially expressed (DE) gene sets predict changes in regulation. Differentially expressed genes are a consequence of a lack of functionality or mis-expression of a particular Transcription Factor, which triggers the downstream mis-expression of its direct target genes and a vast amount of indirectly related genes. Successful motif discovery tools in human are ModuleMiner, PhylCRM/Lever and i-Regulon (Janky *et al*. 2014; Van Loo *et al*. 2008; Warner *et al*. 2008).

In our opinion, the most promising approach for genomic selection is to utilize the information of how genes interact with each other, i.e., pathway analyses or 'gene set analyses' (GSEA). There are several tools for pathway analysis and, broadly, three kinds of functional pathway analysis: over-representation analysis, functional class scoring and pathway topology. Over-representation analysis requires that the input is a list of DE genes, this method evaluates the genes in a specific pathway that show changes in expression, counting the number of DE genes that are in the pathway. Functional class scoring analysis uses the entire data as input, this method follows three steps: first, computes differential expression of individual genes or proteins; second, the gene-level statistics of the genes of a specific pathway are aggregated into a single pathway-level statistic; finally, estimates the statistical significance of this pathway-level statistic. Pathway topology analysis uses the number and type of interactions between gene products, this method is essentially the same as the functional class scoring method but with the difference that the pathway topology analysis uses the additional information of the genes to compute the gene-level statistics.

COMBINING DISPARATE SOURCES OF INFORMATION

Formal integration of information from seemingly disparate sources aims at elucidating the congruency of these sources to further gain biological insight in a manner not possible by each individual source in isolation. The underlying premise is that inaccuracies are less likely to be present when separate data sources corroborate each other. Most applications for combining disparate sources in molecular biology follow one of three general approaches: meta-analysis, graph theory and cluster analysis. Here, we provide a brief description of each approach and when available present and discuss references of relevance to animal breeding and genetics.

Meta-analysis can be seen as an attempt to increase sample size. The objective is to achieve a higher statistical power by aggregating the results from separate studies linked by a common measure such as the effect of a SNP or the abundance of a gene. PRISMA is an organisation that provides guidelines for the systematic reporting of meta-analyses (http://www.prisma-statement.org/index.htm). Many journals endorse the PRISMA guidelines and require their authors to adhere to them. As an example, Pérez-Montarelo *et al.* (2012) undertook a meta-analysis of 20 gene expression studies in porcine spanning 134 experimental conditions on 27 distinct tissues. In an attempt to control the experimental design effects that may contribute to bias, the authors normalised the data by fitting a mixed-model approach that accounted for the disparity in the origin of the studies. With a focus on transcription factor genes and tissue-specific genes, a gene co-expression network was inferred where genes clustered by tissue and tissues clustered by embryonic origin. In another example, and in order to characterise inbreeding depression across species and traits, Leroy (2014) conducted a meta-analysis on 57 studies, 37 phenotypes and seven livestock species. Reported estimates of inbreeding depression were analysed using a multiple regression model that included the effect of study and phenotype. As result, the author reported an average decrease of 0.35% of the mean of a trait per 1% of inbreeding.

Graph-theoretic approaches have the intuitive appeal of network systems where objects (typically genes) are represented by nodes and relationships (typically interactions) are represented by edges. A number of attributes can be overlaid in the visualization schema and the resulting network visualized and explored using a (more or less friendly) software platform such as Cytoscape (www.cytoscape.org). Beyer and May (2003) developed a graph-theoretic algorithm, namely PARTITION, to the partition of individuals into full-sib families. Input to the algorithm is a list of individuals and their genotypes at each locus. For each pair of individuals, a likelihood ratio is calculated from the likelihood of being truly full-sibs over the likelihood that the pair is unrelated. The output is a list of full-sib families in the data set. A second example of graph-theoretic approaches is the work of Balasubramanian *et al.* (2004), who presented an approach for testing the association between multiple sources of functional genomics data, namely the edge permutation and node label permutation tests.

Finally, Bayesian correlated clustering (eg. Kirk *et al.* 2012), and Bayesian consensus clustering (Lock and Dunson 2013) are gaining momentum in the simultaneous integration of information from a wide range of different datasets and data types. In correlated clustering, the allocation of objects (e.g. genes) to clusters in one dataset has an influence on the allocation of genes to clusters in another dataset. Instead, consensus clustering is most commonly used to combine multiple clustering algorithms, or multiple realizations of the same clustering algorithm, on a single dataset.

TOWARDS A BIOLOGY INFORMED BREEDING ECOSYSTEM

The usefulness of sequence or high-density genotyping for genetic prediction is likely to reach a plateau rapidly, when used in isolation. In other words, there is so much redundancy in this kind of data that the likelihood ratio becomes flat when comparing alternative models with varying SNP density. In our opinion, the most promising way to move forward is by embracing the ‘big data’ paradigm. However, contrary to what is normally understood by 'big data', the challenge is not in its size, but rather in its heterogeneity. Very much like internet companies try to make sense of the wide array of information collected by their clients in order to predict their behaviour, animal breeding companies should combine in an optimal way the huge public datasets containing biological information with their own phenotypic and polymorphism data. This is, admittedly, a vague recommendation and there is not, as of today, closed recipes to make the most of this information.

CAUTIONS

Even if very short and incomplete, this review points to the main issue that genomic selection will be facing if external biological information is to be successfully employed: how to weigh in an optimal way the vast diversity of external data that is already available. Our starting hypothesis is that there is not enough information in the data (i.e., in the likelihood) to tell whether a SNP is of sufficient relevance to be included or not in the predictive model and that, therefore, external information will be key to the successful use of sequence data.

REFERENCES

Andersson, L., Archibald, A.L., Bottema, C.D., *et al*. (2015) *Genome Biol*. **16**: 57.

Beyer, J., and May B. (2003) *Mol. Ecol*. **12**: 2243.

Balasubramanian, R., LaFramboise, T., Scholtens, D., *et al*. (2004) *Bioinforma. Oxf. Engl.* **20**: 3353.

Brawand, D., Soumillon, M., Necsulea, A., *et al*. (2011) *Nature* **478**: 343.

Daetwyler, H.D., Capitan, A., Pausch, H., *et al*. (2014) *Nat. Genet.* **46**: 858.

Druet, T., Macleod, I.M., Hayes, B.J. (2014) *Heredity* **112**: 39.

Hayes BJ, MacLeod IM, Daetwyler HD, *et al*. (2014). In *Proc.10th World Cong. Genet. App. Livest. Prod*.

Janky, R., Verfaillie, A., Imrichová, H., *et al*. (2014) *PLoS Comput. Biol*. **10**: e1003731.

Kirk, P., Griffin, J.E., Savage, R.S., *et al*. (2012) *Bioinforma. Oxf. Engl*. **28**: 3290.

Leroy, G. (2014) *Anim. Genet*. **45**: 618.

Liao, X., Bao, H., Meng, Y., *et al*. (2014) *PloS One* **9**: e102868.

Lock, E.F., Dunson, D.B. (2013) *Bioinformatics*. **29**: 2610.

MacLeod, I.M., Hayes, B.J., Goddard, M.E. (2014) *Genetics* **198**: 1671.

Meuwissen, T., Goddard, M. (2010) *Genetics* **185**: 623.

Perez-Enciso, M., Rincón, J., & Legarra, A. (2015). *Genet. Sel. Evol.*, in press

Pérez-Montarelo, D., Hudson, N.J., Fernández, A.I., *et al*. (2012) *PloS One* **7**: e46159.

Schmidt, D., Wilson, M.D., Ballester, B., *et al.* (2010) *Science* **328**: 1036.

Van Loo, P., Aerts, S., Thienpont, B., *et al*. (2008) *Genome Biol*. **9**: R66.

Villar, D., Berthelot, C., Aldridge, S., *et al*. (2015) *Cell*. **160**: 554.

Warner, J.B., Philippakis, A.A., Jaeger, S.A., *et al.* (2008) *Nat. Methods* 5, 347.

Table 1. Selected list of sites containing biological information

Database	Website	Description
Sequence		
GenBank	http://www.ncbi.nlm.nih.gov/entrez	An annotated collection of all publicly available DNA sequences.
EMBL	http://www.ebi.ac.uk/	Framework that provides free access to a range of mainstream sequence analysis applications.
DDBJ	http://www.ddbj.nig.ac.jp/	Primary nucleotide sequence database that provides analytical resources for biological information.
Protein		
SWISS-PROT	http://www.expasy.org/sprot	Swiss-Prot is the section of UniProtKB (central hub of protein knowledge) where the information is manually curated.
PIR	http://pir.georgetown.edu/	Resource that provides protein databases and analysis tools to support research on molecular evolution, functional genomics and computational biology.
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop	Database that provides a detailed and comprehensive description of the relationships of all known proteins structures.
Genomic		
Entrez biosystems	http://www.ncbi.nlm.nih.gov/biosystems/	Database providing integrated access to biological systems and their component genes, proteins, and small molecules, as well as literature describing those biosystems and other related data.
Entrez Genomes	http://www.ncbi.nlm.nih.gov/entrez	Database that contains sequence and map data from the whole genomes of over 1000 organisms.
KEGG	http://www.genome.ad.jp/kegg	Database of biological systems that integrates genomic, chemical and systemic functional information.
Organism-specific		
AnimalQTLdb	http://www.animalgenome.org/cgi-bin/QTLdb/	Contains reported QTL in livestock
FlyBase	http://flybase.bio.indiana.edu/	Database of genetic and genomic data for the insect family *Drosophilidae*.

OMIM	http://www.ncbi.nlm.nih.gov/Omim	Knowledgebase of human genes and phenotypes.
Transcription factor binding		
AnimalTFDB	http://www.bioguo.org/AnimalTFDB/	TF database specialized in livestock
TRANSFAC	http://transfac.gbf.de/	
DBD	http://www.transcriptionfactor.org/	Database of predicted transcription factors in completely sequenced genomes and their sequence specific DNA-binding domain families.
Epigenetic databases		
Epigenomics	http://www.ncbi.nlm.nih.gov/epigenomics	Resource for whole-genome epigenetic data sets.
MethDB	http://www.methdb.de/	Database for DNA methylation and environmental epigenetic effects.
The Histone Database	http://genome.nhgri.nih.gov/histones/	Resource for histones and histone fold-containing proteins.
CREMOFAC	http://www.jncasr.ac.in/cremofac/	Database dedicated for chromatin-remodeling factors.
Biochemical databases		
ENZYME	http://www.expasy.org/enzyme	Repository of information relative to the nomenclature of enzimes.
BRENDA	http://www.brenda-enzymes.org/	Database on functional and molecular information of enzymes.
AAindex	http://www.genome.ad.jp/dbget/aaindex.html	Database of phyisicochemical and biochemical properties of amino acids.