Integrating Data from Different Marker Panels in Human Genetics

T. Druet
Unit of Animal Genomics, Faculty of Veterinary Medicine and Centre for Biomedical Integrative Genoproteomics, University of Liège (B43), Belgium

Abstract

Until recently, different commercial arrays with 300,000 to one million SNPs have been used for studies in human genetics. Samples from genome-wide association studies have been genotyped on different genotyping platforms, while much more markers were used for reference panels. To increase power of association studies, different data sets are combined by imputing the missing markers. Imputation techniques have proven to be very efficient and allelic imputation error rates were estimated to range between 0.5 and 2.5 %, depending on the study. Although markers densities are lower, these techniques can be applied in cattle too. Thanks to strong familial information and larger samples, low imputation errors rates (below one percent) can be achieved. Imputation error rates are directly influenced by marker density and genetic relationships between target and reference individuals. Even with arrays of 3000 SNPs, imputation can yield reasonable marker predictions. As in human genetics, the development of publicly available databases of influential animals genotyped at high density or even re-sequenced would be beneficial to the field.

1. Introduction

Nowadays different marker panels are available in cattle, such as the Affymetrix 10K SNP Bovine array or the Illumina Bovine SNP50TM chip. In addition, it is possible to design custom marker panels such as the 60K Illumina chip used by CRV and described in Charlier *et al.* (2008). Commercial genotyping arrays with more than 600,000 SNPs are under development (e.g., Illumina BovineHD). It is also expected that animal breeders will use small chip panels (e.g., Illumina Bovine3K) as suggested by Habier *et al.* (2009). Therefore, it is likely that cattle breeders will end up with their population genotyped on different marker panels and will be in need for solutions to combine all the data.

In human genetics, different marker panels are also used. Genotyping technologies are in constant evolution. In the recent past, these panels ranged from approximately 300,000 to more than one million SNPs.

The present paper briefly describes how genotypes from different marker panels are combined in human genetics and discusses whether the same techniques can be applied in cattle.

2. Marker Panels in Human Genetics

Commercial genotyping arrays with approximately 300,000 to 550,000 SNPs proposed by Affymetrix, Illumina or Perlegen are commonly used in genome wide association studies (**GWAS**). Arrays with more than one million SNPs are already available and arrays with up to 5 millions SNPs are projected.

Individuals from the HapMap population (The International HapMap Consortium, 2005) have already been genotyped for 3.1 million SNPs. In the 1000 genomes project, more than 1000 individuals are resequenced (www.1000genomes.org), which is the most complete genotyping technology.

3. Benefits of Combining Data from Different Marker Panels

GWAS for complex traits, such as risk of coronary disease, obesity, schizophrenia, height, Crohn's disease and many others (see for instance the WTCCC), have yielded relatively disappointing results since the identified risk loci explain only a small fraction of the genetic variance associated to the trait of interest. In order to increase the power of these association studies, several research teams combined their data in a meta-analysis study (e.g., Barrett *et al.*, 2008; Willer *et al.*, 2008). Hence, Barrett *et al.* (2008) estimated that such combination allowed them to detect effects below 1.2 odds ratio that were unlikely to be detected in the original designs taken separately. However, samples are not always genotyped on the same marker array, which makes potential meta-analyses more difficult to implement. Another way to increase the power of GWAS is to supplement the sample under study with HapMap genotypes (or genotypes from any other reference panel). In both cases, the most popular method used for combining these data is to predict the missing markers, which is termed "imputation".

4. Marker Imputation

4.1. Principle

Marker imputation consists in locally matching haplotypes of individuals genotyped with a given array with reference sequences obtained from individuals genotyped at higher density. The missing markers of the individuals can then be predicted by using the corresponding markers of the matching reference sequence. Generally, imputation is based only on linkage disequilibrium and reference sequences are obtained without using individuals genotyped on the "lower" density arrays.

4.2. Methods for marker imputation

Imputation methods often rely on haplotyping methods. The most popular methods are IMPUTE (Marchini *et al*., 2007), MACH (Li *et al*., 2006), fastPHASE (Scheet and Stephens, 2006) or Beagle (Browning and Browning, 2007). Each of these approaches can be described with a Hidden Markov Model. IMPUTE uses haplotypes of HapMap individuals as reference sequences. FastPHASE uses these individuals to determine a set of K (generally between 10 and 20) ancestral haplotypes as reference sequences. Beagle constructs a tree with all the haplotypes of the reference panel and then summarizes it in a directed acyclic graph by joining nodes of the tree. Beagle is faster than IMPUTE or fastPHASE. For a review of imputation methods, see Browning (2008) for instance.

4.3. Accuracy

Some examples of accuracies obtained with individuals genotyped on commercial arrays can be found in Browning and Browning (2007) or Willer *et al.* (2008). Browning and Browning (2007) compared Beagle and fastPHASE and obtained allelic imputation error rates below 1% for individuals genotyped on an Affymetrix 500K array. Willer *et al.* (2008), in a meta-analysis studying coronary artery disease, estimated that the allelic imputation error rate was equal to 1.46 for individuals genotyped on the Affymetrix 500K array and 2.14 for those genotyped on the Illumina HumanHap300 array.

4.4. Increased power with imputed data and cost effective genotyping

Anderson *et al.* (2008) or Spencer *et al.* (2009) demonstrated that using, in addition to the genotyped markers, imputed genotypes (based on the HapMap panel) increased the power of GWAS. For instance, with Illumina HumanHap300 and Affymetrix 500K arrays, power increased from 0.392 and 0.363 to 0.467 and 0.450, respectively (Spencer *et al.*, 2009). The same authors concluded that the most cost effective genotyping design (measured as the highest power at constant cost) was achieved by using the Illumina HumanHap 300 array rather than higher density genotyping arrays.

4.5 New variants identified

In several studies, new variants were identified using the imputation technique. These variants had not been identified with traditional approaches in initial experiments. For instance, Barret *et al.* (2008) in Crohn's Disease or Purcell *et al.* (2009) in schizophrenia and bipolar disorder detected new associations in large meta-analyses. Further, the use of imputed markers from the HapMap data set led to identifying new variants for type 2 diabetes in Zeggini *et al.* (2008).

4.6 Reduction of computational costs

Imputation methods are computationally expensive, depending on the number of individuals and markers involved. Even if software optimizations can sometimes reduce the computational burden, strategies are still needed to save CPU time. Hence, splitting chromosomes in pieces allows to reduce total computation time but at the expense of using more processors. Moreover, this approach is not optimal for all positions along the chromosome.

One way to save computational and economical cost applied in human genetics is the development of public reference databases, such as the HapMap project or the 1000 genomes project. Thanks to these databases, reference panels genotyped at high density (or re-sequenced) are available to everyone. In addition to the raw genotypes, processed data is also shared: phased data, linkage disequilibrium, recombination maps, imputation model parameters, GWAS results and sometimes phenotypes.

5. Use of Imputation in Dairy Cattle

5.1. Linkage disequilibrium and marker density

Comparison of LD patterns in human (e.g., Jakobsson *et al.*, 2008) and cattle (e.g., Gautier *et al.*, 2007) show that, at equal marker density, imputation can be performed in cattle as efficiently as in human. However, genotyping arrays used in human rely on at least 300,000 SNPs whereas cattle, arrays are currently limited to 50,000 SNPs. In addition, low density chips with approximately 3000 SNPs are planned to be used in cattle (e.g., the Illumina Bovine3K). Therefore, imputation in cattle will rely on lower density maps for a while. On the other hand, other factors might favor efficient imputation in cattle, such as the availability of strong familial information and the huge sizes of some of the genotyped populations.

5.2. Application of imputation in cattle

Hayes *et al.* (2009) applied imputation techniques with fastPHASE (Scheet and Stephens, 2006) on bulls genotyped with a 50K marker panel. With a few missing genotypes, they estimated that the allelic imputation error rate was approximately equal to 1.3%.

T. Druet, C. Schrooten and A.P.W. de Roos, A.P.W. (in preparation) applied Beagle (Browning and Browning, 2007) and DAGPHASE (Druet and Georges, 2010) to carry out the imputation for the EuroGenomics project. In the preceding feasibility study, they divided a bovine 50K array into two chips of approximately 27.5K SNPs. Then, the 1000 individuals with the largest number of descendants were genotyped on all the markers (i.e., were doubly genotyped) as a reference panel. To evaluate the imputation efficiency, two designs were tested. In the first design, all the 3738 remaining animals were genotyped on the same chip whereas in the second design, 2351 and 1387 animals were genotyped on each chip. Overall allelic imputed errors rates were equal to 0.65% and 0.50% with the first and second design, respectively. A number of parameters were shown to affect the allelic imputation error rate, including the size of the reference panel (number of animals genotyped on all markers), marker density and genetic relationships between target and reference individuals.

Figure 1 shows the effect of marker density on the allelic imputation error rate. Marker density was measured as the number of genotyped markers in the Mb surrounding the imputed marker. At low marker densities, imputation error rates decrease abruptly with the number of markers. For instance, error rates were above 3% with zero or one marker in the Mb surrounding the imputed marker and below 1% with so few as 5 markers. At the higher marker densities available today (10 to 20 markers per Mb), allelic imputation error rates ranged from 0.66% to 0.50%, while they were still decreasing, albeit slowly, at even higher densities, thus suggesting that upcoming high-density marker panels will effectively improve imputation.

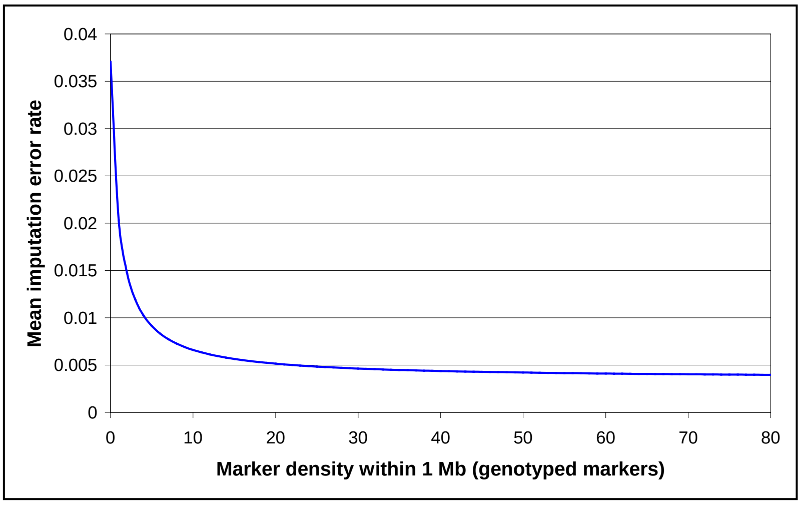

Figure 1. Effect of marker density on the allelic imputation error rate.

In addition to the effect of marker density, the influence of genetic relationships between target and reference individuals was also studied. This relationship for a given pair of individual was scored as the expected proportion of the genome inherited from the reference individuals by the target individual. For instance, this score is equal to 0.5 when a single parent is genotyped on all markers and 1.0 when both parents belong to reference individuals. The relationship between this genetic score and the allelic imputation error rate is described in Table 1. As expected, the imputation efficiency increases with the score. When both parents were genotyped, the imputation error rate was below 0.15%, with imputation relying only on linkage. For other haplotypes, LD information is used when the parent is not genotyped. Finally, the allelic imputation error rates were also estimated with a low density chip containing 3000 SNPs. In that case, imputation error rates increased above 5% for scores below 0.75. For most animals, errors rates ranged between 3 and 4% whereas they dropped to 0.5% when both parents were genotyped.

Table 1. Allelic imputation error rate as a function of the relationship "score".

Score	Imputation error rate
< 0.75	1.16 %
0.75 ≤ . <0.90	0.91 %
0.90 ≤ . <0.95	0.79 %
0.95 ≤ . <1.00	0.70 %
1.00	0.14 %

5.3. Computational issues

With the increasing number of genotyped animals and/or markers, important computational issues arise. Several options can be applied to reduce the computational burden. First, the imputation step can be dropped or reduced by constructing genomic relationship matrices relying only on genotyped SNPs (Legarra *et al*., 2009), by using only 50K SNPs for genomic selection or by focusing only on a subset of genomic regions. CPU requirements can also be reduced by applying strategies similar to variance components estimation, i.e., estimating model parameters less frequently (not at each imputation) and only on a subset of the genotyped individuals (although parameters are estimated more precisely when using all the data). In addition, imputation can be carried out sequentially. To impute newly genotyped animals, only genotyped relatives and model parameters are required. Therefore, the imputation does not need to be performed on the complete data set. Finally, public reference databases (as those implemented in human genetics) could save both genotyping costs and CPU time. Indeed, for imputation (and other operations too), a reference panel genotyped at a higher density is required. Sharing such a reference panel for one breed would certainly lead to big savings in comparison to using one different reference panel in each country. Another advantage of this approach is that parameter estimation could be performed once for all on the common reference panel and shared among partners. Some pre-processed data, such as phased data, might also be shared.

6. Conclusions

In human genetics, individuals genotyped on different marker arrays are combined into meta-analyses in order to increase the power of association studies. Further, reference panels genotyped at high density are jointly analyzed with samples genotyped on commercial arrays using imputation techniques. These techniques can also be successfully applied in cattle where large data sets with strong familial information are available. Imputation in cattle might also benefit from shared public reference databases with important individuals genotyped at high density.

Acknowledgements

Tom Druet is Research Associate from the Fonds de la Recherche Scientifique – FNRS. I acknowledge Zhiyan Zhang (Unit of Animal Genomics, University of Liège), Sander de Roos and Chris Schrooten (CRV, The Netherlands) and the partners of the EuroGenomics project for their collaboration in imputation studies.

References

Anderson, C.A., Pettersson, F.H., Barrett, J.C., Zhuang, J.J., Ragoussis, J., Cardon, L.R. & Morris, A.P. 2008. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet. 83(1),* 112-119.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., Bitton, A., Dassopoulos, T., Datta, L.W., Green, T., Griffiths, A.M., Kistner, E.O., Murtha, M.T., Regueiro, M.D., Rotter, J.I., Schumm, L.P., Steinhart, A.H., Targan, S.R., Xavier, R.J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.P., de Vos, M., Vermeire, S., Louis, E., Cardon, L.R., Anderson, C.A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N.J., Onnie, C.M., Fisher, S.A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C.G., Parkes, M., Georges, M. & Daly, M.J. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet. 40(8),* 955-962.

Browning, S.R. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet. 124(5),* 439-450.

Browning, S.R. & Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet. 81(5),* 1084-1097.

Charlier, C., Coppieters, W., Rollin, F., Desmecht, D., Agerholm, J.S., Cambisano, N., Carta, E., Dardano, S., Dive, M., Fasquelle, C., Frennet, J.C., Hanset, R., Hubin, X., Jorgensen, C., Karim, L., Kent, M., Harvey, K., Pearce, B.R., Simon, P., Tama, N., Nie, H., Vandeputte, S., Lien, S., Longeri, M., Fredholm, M., Harvey, R.J. & Georges, M. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat. Genet. 40(4),* 449-454.

The International HapMap Consortium, 2005. A haplotype map of the human genome. *Nature 437(7063),* 1299-1320.

Druet, T. & Georges, M. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fie mapping. *Genetics 184 (3).*

Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Foglio, M., Grohs, C., Boland, A., Garnier, J.G., Boichard, D., Lathrop, G.M., Gut, I.G. & Eggen, A. 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics 177(2),* 1059-1070.

Habier, D., Fernando, R.L. & Dekkers, J.C. 2009. Genomic Selection Using Low-density Marker Panels. *Genetics.*

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. 2009. Invited review:

Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci. 92(2),* 433-443.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., Bras, J.M., Schymick, J.C., Hernandez, D.G., Traynor, B.J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H.M., Hardy, J.A., Rosenberg, N.A. & Singleton, A.B. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature 451(7181),* 998-1003.

Legarra, A., Aguilar, I. & Misztal, I. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci. 92,* 4656-4663.

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet. 39(7),* 906-913.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F. & Sklar, P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature 460(7256),* 748-752.

Scheet, P. & Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet. 78(4),* 629-644.

Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet 5(5)*:e1000477.

The International HapMap Consortium, 2005. A haplotype map of the human genome. Nature 437(7063):1299-1320.

Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, S. C. Heath, N. J. Timpson, S. S. Najjar, H. M. Stringham, J. Strait, W. L. Duren, A. Maschio, F. Busonero, A. Mulas, G. Albai, A. J. Swift, M. A. Morken, N. Narisu, D. Bennett, S. Parish, H. Shen, P. Galan, P. Meneton, S. Hercberg, D. Zelenika, W. M. Chen, Y. Li, L. J. Scott, P. A. Scheet, J. Sundvall, R. M. Watanabe, R. Nagaraja, S. Ebrahim, D. A. Lawlor, Y. Ben-Shlomo, G. Davey-Smith, A. R. Shuldiner, R. Collins, R. N. Bergman, M. Uda, J. Tuomilehto, A. Cao, F. S. Collins, E. Lakatta, G. M. Lathrop, M. Boehnke, D. Schlessinger, K. L. Mohlke, and G. R. Abecasis. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet. 40(2),* 161-169.

Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen, K. Ardlie, K. B. Bostrom, R. N. Bergman, L. L. Bonnycastle, K. Borch-Johnsen, N. P. Burtt, H. Chen, P. S. Chines, M. J. Daly, P. Deodhar, C. J. Ding, A. S. Doney, W. L. Duren, K. S. Elliott, M. R. Erdos, T. M. Frayling, R. M. Freathy, L. Gianniny, H. Grallert, N. Grarup, C. J. Groves, C. Guiducci, T. Hansen, C. Herder, G. A. Hitman, T. E. Hughes, B. Isomaa, A. U. Jackson, T. Jorgensen, A. Kong, K. Kubalanza, F. G. Kuruvilla, J. Kuusisto, C. Langenberg, H. Lango, T. Lauritzen, Y. Li, C. M. Lindgren, V. Lyssenko, A. F. Marvelle, C. Meisinger, K. Midthjell, K. L. Mohlke, M. A. Morken, A. D. Morris, N. Narisu, P. Nilsson, K. R. Owen, C. N. Palmer, F. Payne, J. R. Perry, E. Pettersen, C. Platou, I. Prokopenko, L. Qi, L. Qin, N. W. Rayner, M. Rees, J. J. Roix, A. Sandbaek, B. Shields, M. Sjogren, V. Steinthorsdottir, H. M. Stringham, A. J. Swift, G. Thorleifsson, U. Thorsteinsdottir, N. J. Timpson, T. Tuomi, J. Tuomilehto, M. Walker, R. M. Watanabe, M. N. Weedon, C. J. Willer, T. Illig, K. Hveem, F. B. Hu, M. Laakso, K. Stefansson, O. Pedersen, N. J. Wareham, I. Barroso, A. T. Hattersley, F. S. Collins, L. Groop, M. I. McCarthy, M. Boehnke, and D. Altshuler. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet. 40(5),* 638-645.