Summary Focus Meeting Imputation

David van Enckevort

Introduction

Genome wide imputation of untyped variants within the genome has been one of the biggest drives behind the recent "gold rush" of identifying genes for various complex traits in human, animal and plant genetics. One major advantage of imputation is making (human) GWAS studies comparable even though different genotyping procedures have been applied, and the imputed variants enhance the genomic scans as the extra information provides a much clearer picture of the possible associations. Imputation has become standard practice, and genotypes of imputed variants are considered to be valid measures. In agricultural species, imputation plays a critical role in cost-saving genotyping strategies in genomic prediction of large pools of selection candidates.
However, behind the advantages of imputation lies a complex field of rapid developing methods and decisions, which may affect the outcome of association studies. Progress in quality of the imputation as well as improving the calculation speed of imputing millions of variants are basically made monthly by statisticians, bio-technicians and geneticists. However, the numbers of variants, subjects in the datasets and reference sets are increasing with the same speed. It leads to a demanding field, where researchers continuously have to keep updated with developments.
During the day four presentations were planned:

1. Marco Bink (WU): Genotype Imputation in Plants;
2. Mario Calus (WU): Genotype Imputation in livestock;
3. Stefan Böhringer (LUMC): Technical aspects of reproducible imputation pipelines;
4. Jouke-Jan Hottinga (VU): Human DNA Imputation Progression & Pitfalls.

Slides of the presentations are attached to this document. Due to unexpected circumstance the presentation of Jouke-Jan Hottinga had to be cancelled.

Summary and conclusions

In general imputation is done with different goals:

I. Impute randomly missing genotypic data;
II. Impute genotypic data for alignment of different SNP arrays;
III. Impute genotypic data from low-density SNP array to high-density SNP array;
IV. Impute genotypic data from low coverage sequencing data (including genotyping by sequencing).

In livestock and plants a main purpose is to identify mutations for genomic selection to improve favourable traits (e.g. improved resistance to common crop diseases; resistance to drought; taste or nutritional value), while in human genetics the main purpose is to compose large cohorts for Genome-Wide Association Studies (GWAS) for disease genotypes and complex traits.
Plants have a huge genomic complexity compared to human and livestock due to variance in length, ploidy-level and structure (duplication, repeats) and there is heterogeneity in species. For most species there are scarcely markers available, the

situation is better in livestock and for human genetics we have a good reference genome. In plants and livestock the main focus is on iii) and iv) and there are scarcely examples for ii) to increase accuracy and reduce costs.
There are different methodological challenges in imputation when combining data from different partly overlapping SNP arrays, which are mainly relevant in GWAS and meta-analysis in human genetics. Both in livestock breeding and plant genetics demonstrate good results in using imputation to improve accuracy when combining low-density SNP arrays with high-density SNP arrays, which provides a good balance between costs and quality.
Imputation software and pipelines have much improved over the last 5 years and standard protocols for imputation has been established in large research consortia. Even though studies deal with ever increasing population sizes and reference sets the improvement in tools and protocols and tools to manage your data keep the process maintainable. However there is still a gap in the reproducibility of imputation caused by the complexity of the pipelines and difficulties in version management of tools and datasets. There is room for a formalisation of pipelines and frameworks for running pipelines have been developed in several locations.
There is good consensus between the three different domains on methodology and the software that is being used. Problems around management of the data and tool versions are shared between the domains. There is a proposed solution in the formalisation of the pipelines, which is part of the broader need to document provenance and develop data stewardship.