In silico candidate gene mining in livestock species

A. Stella, A. Montironi, F. Panzitta, G. Gandini & P. J. Boettcher

In silico candidate gene mining in livestock species

A. Stella[1], A. Montironi[1,2], F. Panzitta[1],
G. Gandini[2], P. J. Boettcher[3,4]

[1] Parco Tecnologico Padano. Lodi, Italy
[2] Dipartimento di Scienze e Tecnologie Veterinarie per la Sicurezza Alimentare.
Università di Milano, Italy
[3] Istituto di Biologia e Biotecnologia Agrarie. Consiglio Nazionale delle Ricerche. Milano, Italy
[4] International Atomic Energy Agency. Wien, Austria

Corresponding author: Alessandra Stella. Parco Tecnologico Padano. Via Einstein, Località Cascina Codazza, 26900 Lodi, Italy - Tel. +39 0371 4662690 - Fax: +39 0371 4662419 - Email: alessandra.stella@tecnoparco.org

ABSTRACT

The identification of genes involved in livestock production and disease is a challenge due to the multi-genic, multi-factorial nature of the traits and the complexity of integration of information from different studies. Genome-wide techniques such as microarray analysis, SAGE, linkage analysis and linkage disequilibrium analysis have been extensively used in livestock and have often identified a large number of candidate genes. Selection of the most probable candidate genes for further empirical analysis remains a challenge. Novel extensive biological databases (DB) provide an opportunity for candidate gene mining. Bioinformatic methods and tools to prioritize candidate genes underlying pathways or diseases have been presented mostly for application to human disease candidate gene search. These computational methods employ data from a variety of sources to identify the most likely candidate genes from genes sets. The objectives of the study were: 1. to test a set of existing gene prioritization computational methods on real and simulated livestock traits, namely mastitis susceptibility in cattle, production in cattle, litter size in swine, and tick resistance in cattle; 2. to develop a novel method for candidate prioritization that better suits the characteristics of genomic information of livestock species (lower level of annotation, different experimental designs, etc.). The algorithm performs distinct prioritizations from multiple heterogeneous data sources, which are then integrated into one global ranking using order statistics. Information about a trait or pathway is recorded, ordered and stored from a set of known genes using multiple data sources. Then, the candidate genes are ranked based on similarity with the training properties obtained in the first step, resulting in one prioritized list for each data source. Data for linkage and association analysis, and expression analysis (2 microarray studies) were simulated for a complex trait assigning the largest effects to known, well described genes in the Gene Ontology (GO), InterPro, MEDLINE, and Kegg databases. All the other simulated major genes were assigned to genes described in one of the databases for livestock species. Real data were selected from the literature, mainly from large QTL studies, or obtained from collaborators. Software used for the analysis were: Suspect (http://www.genetics.med.ed.ac.uk/software/prospectr.php), Endevour and COeXpress (http://coxpress.sf.net). Results using simulated data showed that where annotation is missing, the accuracy of the considered algorithms decreased drastically. A new prioritization method applied to simulated data correctly ranked candidate gene only when QTL information had a high level of accuracy. Further work is needed in defining methods of weighting QTL data information.