Gene prioritization for livestock diseases by data integration

Li Jiang,1 Peter Sørensen,1 Bo Thomsen,1 Stefan M. Edwards,1 Axel Skarman,1 Christine M. Røntved,2 Mogens S. Lund,1 and Christopher T. Workman3

Departments of 1Molecular Biology and Genetics and 2Animal Science, Aarhus University, Tjele; and 3Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark

Submitted 17 March 2011; accepted in final form 10 January 2012

Mastitis is an inflammatory reaction in the bovine mammary gland caused due to invasion and colonization by diverse pathogens including contagious and environmental bacteria such as Streptococcus spp., Staphylococcus spp., and coliform bacteria (3). Initiation and resolution of inflammation are controlled by a complex network of biological substances that belong to the homeostatic armamentarium of the host (6). Mastitis is the most common infectious disease in dairy cattle worldwide, resulting in reduced animal welfare and enormous economic losses. We have successfully used linkage and linkage disequilibrium analyses (23, 39) as well as genome-wide association analysis to identify several quantitative trait loci (QTL) affecting the resistance to clinical and subclinical mastitis (22, 33). According to the sequenced reference bovine genome (8), the sizes of these genomic regions were typically in the range of megabases, each containing a plethora of genes of which some play a critical role in mastitis resistance either as single genes or in combination, for example as interacting partners in protein complexes. Further investigations showed that these QTL affected the levels of genetic resistance specifically against the most common mastitis pathogens, including Staphylococcus aureus and Escherichia coli (38). To complement these studies and to gain further insight into the genetic background for mastitis-related traits, we also applied genome-wide expression analyses, which identified global changes in expression levels of genes and revealed genetic pathways that are associated with bovine mastitis (5, 17). Despite the transcriptome-revealed variation of gene regulation in response to infection, the spatial-temporal profile cannot depict the general pattern of bovine gene with respect to disease relevance. The key challenge lying ahead is to prioritize and select the most probable candidate gene(s) for in-depth molecular analysis from the long list of differentially expressed genes caused by mastitis infection or among the hundreds of genes located in the mastitis-associated QTL regions.

Prioritization of candidate genes needs cumulative information from different levels of a biological system. Biomedical records (e.g., PubMed articles and OMIM entries) in life science provide an immense amount of existing knowledge on phenotypes and gene phenotype relationships derived from studies in various biological conditions. A global examination of these textual data will thus establish the phenotypic profile of genes with respect to their connection to disease biology.

Network-based disease gene prioritization relies on the observation that genes that are functionally linked in pathways or protein complexes are often associated with identical or similar phenotypes (7, 9, 10, 18, 20, 42–45). Thus, when one or more proteins in a given protein complex are already implicated in a disease of interest, then the remaining network members im-
Gene prioritization by data integration

Network-based gene prioritization typically integrates information from disease phenotypes, gene-associated phenotypes, and protein-protein interactions (PPI) and has been substantially studied, validated, and implemented in several bioinformatic approaches (7, 9, 10, 18, 20, 42–45). We reasoned that network-based disease gene prioritization would also be beneficial for prioritizations of candidate genes for complex diseases in livestock species.

The phenome profiling is based on existing knowledge, which means that the current incomplete characterization of genes in terms of function and correlation between genes and diseases will allow only part of genome to be prioritized. However, the limitation in gene phenotype relationships and protein associations could be alleviated by using ortholog information, which allows us to take advantage of the richly annotated biomedical databases of human and mouse.

Therefore, the aims of this study are to 1) implement a network-based disease gene prioritization that integrates protein networks and biomedical records from different species for identifying and prioritizing genes according to their relevance to a complex disease, and 2) to use this approach to identify and rank candidate genes involved in bovine mastitis.

Materials and Methods

The study was designed to implement and apply a network-based disease gene prioritization approach for ranking bovine genes according to their relevance to the mastitis. The starting point of the approach is a list of candidate genes that potentially underlie the disease, for example genes that are differentially expressed, genes located in a specific genomic region, or even all known genes present in the genome. A network-based gene prioritization approach that integrates information from disease phenotypes, gene-associated phenotypes, and PPI was implemented. Two distinct types of gene-associated phenotypes were considered, including those from text mining of biomedical text records (denoted phenome profiling) and those from genome-wide expression profiles (denoted transcriptional profiling). Ortholog mapping was used to take advantage of the richly annotated biomedical databases of other species such as human and mouse, thereby circumventing the limitation in gene-associated phenotypes and PPI in livestock species. The approach is illustrated using phenome profiling in Fig. 1, and a more detailed description of the key steps is given below.

Network-based gene prioritization approach. The prioritization approach works as follows. For each candidate gene a candidate complex is determined from PPI data. Two distinct types of gene-associated phenotypes are the linked to candidate gene complex. First, gene-associated phenotypes were obtained from text mining of biomedical text records. Each gene in the complex is then linked to its biomedical text records in the OMIM (12), PubMed, or GeneRIF (26) databases. Text mining is used to convert the biomedical text into a vocabulary list based on Unified Medical Language Systems (UMLS) (21). For each gene the result is a vector of weighted counts of occurrence of each of the UMLS terms present in the biomedical text record. This vector defines a standardized gene associated phenotype. A vector of standardized disease phenotype is determined in a similar way. In both cases the terms used can be limited to specific vocabularies by applying different vocabulary control filters [e.g., Medical Subject Heading (MeSH), Gene Ontology (GO)]. A disease relevance score for each candidate gene is computed from the pair-wise semantic similarities between the disease phenotype and each of the gene associated phenotypes in the complex. Second, gene-associated phenotypes were also obtained from genome-wide expression profiles. The disease relevance score for each candidate gene is computed based on the logarithms of fold changes (logFC) of genes in the complex. A random-set scoring model (27) is used to calculate the disease relevance score. The random-set scoring model gives a z-score per gene (complex) representing an overall enrichment signal for the association of the candidate gene (complex) to the disease. The z-score is used to determine whether the gene is associated to the disease or not, and it is used for ranking comparing genes for their disease relevance.

Websites. More information regarding this network-based approach using phenome can be found on the project website at https://djfextranet.agrsci.dk/sites/txtphenome/public/Pages/front.aspx. This site contains the R-package txtPhenome and data packages containing processed data from OMIM, GeneRIF, and PubMed. txtPhenome contains functions for dealing with the data in the data packages.

Disease phenotypes for phenome profiling. The phenotypic profile of the disease, referred to as the disease phenotype, can be obtained from a textual description of the disease characteristics such as pathogenesis and clinical features. Ideally this should be based on experts reviewing information retrieved from the published literature such as those of the OMIM database for human diseases. Here, to rank bovine genes according to their potential role in mastitis, we used a disease phenotype description from Wikipedia as an analog to authoritative peer-reviewed, reference-based resources (e.g., OMIM entry for human disorders). Wikipedia is a user-edited encyclopedia based on the principle of collective intelligence, which has been shown to equal the Encyclopedia Britannica in accuracy on scientific entries (11). The specific wiki entry used in the study: (http://en.wikipedia.org/wiki/index.php?title=Mastitis&oldid=263884847).

Gene-associated phenotypes from biomedical text data. Gene-associated phenotypes were obtained from text mining of biomedical text data (e.g., OMIM, GeneRIF, PubMed records). Biomedical text data provide a rich compendium for genes or their protein products in relation to observed biological phenomena and known mechanisms of biological systems. The biomedical text records linked to Entrez genes were from three repositories including 1) the text field of OMIM records linked to human genes, 2) the titles and abstracts of PubMed articles linked to human genes, and 3) the text field of GeneRIF records linked to human, mouse, or cattle genes. The links between species-specific Entrez gene identifiers and the biomedical text records were available from the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/gene/DATA/).

Gene-associated phenotypes from genome-wide expression profiling. Gene-associated phenotypes were also obtained from genome-wide expression profiles. The expression profiles from dairy cattle in response to E. coli infection has been described elsewhere (5). In brief, 16 healthy primiparous Danish Holstein-Friesian cows were challenged intramammarily with E. coli (k2bb2) (20–40 CFU/ml) 4–6 wk after parturition in one quarter. Udder biopsies were collected from the infected quarter, and a healthy control quarter of each cow at 24 h (acute stage) after the E. coli challenge. In total 21 samples were used (12 infected quarters and 9 control quarters). The expression profiles were measured using the Affymetrix Bovine Genome array (Affymetrix, Santa Clara, CA). Descriptions of the microarray experiment and data are available from NCBI’s Gene Expression Omnibus data repository through accession number GSE24217. The annotation and analysis were performed using R (version 2.10.0). The annotation of the probes was provided by Affymetrix and also retrieved using the BiomaRt package (version 2.2.0). Normalization of the arrays was performed using the robust multiarray analysis algorithm (15). The contrast of infected versus control tissue was analyzed. Differential expression of each gene was assessed using linear modeling and empirical Bayes methods (37), which were implemented using the R package LIMMA (version 2.18.2).

Candidate complexes. The candidate complex for each gene was determined from PPI data obtained from the STRING database (version 8.1). STRING is a database and web resource dedicated to PPI, including both physical and functional interactions (16, 41).
Ensembl protein encoded by the candidate gene was used to retrieve first-order protein interaction partners. Since PPI often are evolutionarily conserved (2, 30), interaction data from human or mouse were also used to define orthologous candidate complexes in livestock species. This process involves first mapping the candidate gene into its ortholog in human or mouse and then collecting the protein interaction partners for the corresponding species. The complexes contain the protein encoded by a candidate gene and the first-order neighbors in the network that associate directly with this protein. Each pair of protein association from STRING was annotated with a confidence score ranging from 150 (low credibility) to 999 (high credibility). Protein complexes were retrieved with a filter that each association pair were required to have the confidence score at least 900.

**Ortholog information.** Ortholog mapping and conversion between Entrez gene and Ensembl protein identifiers was based on the following BioConductor annotation data packages “org.Bt.e.g..db_2.3.5,” “org.Hs.e.g..db_2.3.6,” “org.Mm.e.g..db_2.3.6,” “hom.Hs.inp.db_2.3.2,” and “hom.Mm.inp.db_2.3.2.” Bovine Entrez gene identifiers were converted into cattle Ensembl protein identifiers that were used for retrieving cattle Ensembl protein identifiers that were used for retrieving cattle candidate complexes from the STRING database. Alternatively, bovine Ensembl protein identifiers were mapped into the orthologous human (mouse) Ensembl protein identifiers, which then were used to collect the corresponding human (mouse) protein complexes. All bovine, human, or mouse Ensembl protein identifiers in the complexes were converted back to their corresponding Entrez gene identifier, which then were linked to various sources of biomedical records.

**Text mining of biomedical records.** The text field of the disease description of mastitis and the text field of the biomedical records from OMIM, PubMed, or GeneRIF were passed through the text-mining program MMTx (29) (version V2.4.C). Each document was mapped into to a set of UMLS (release version: 2009AA) concepts/terms. The frequencies of concepts within the document and occurrences of concepts across documents were calculated. All the concepts were then filtered using a specific vocabulary such that only MeSH terms remained. Finally the term frequency-inverse document fre-
quency (34) was used to weight all the terms in each document. The semantic similarity was computed as the cosine coefficient between the weighted term occurrences in a disease-associated phenotype and a gene-associated phenotype according to the Vector Space Model (35, 36). Since the phenotypic vectors of genes are defined by the type of biomedical records (human OMIM, human PubMed, and GeneRIF of human, mouse, and cattle) five sets of semantic similarities between the disease phenotype and the phenotypic profile of each genes were produced.

**Disease relevance score.** A random-set scoring model (27) was used to calculate the disease relevance score. In the case of phenome profiling this is computed from the pair-wise semantic similarities between the disease phenotype and each of the gene associated phenotypes in the complex. In the case of transcriptome profiling it is based on a measure of differential expression such as the logFC. The random-set scoring model measures the enrichment signal of the association between the disease phenotype and the set of gene-associated phenotypes linked to the candidate gene complex. The association is represented by a set of element-level scores (log-fold changes or semantic similarities) denoted \( S_e \) that contains \( m \) elements, either the number of transcriptional profiled genes in the candidate complex or the number of biomedical records linked to the genes in the candidate complex. The unstandardized enrichment signal is defined as (27)

\[
\tilde{X} = \frac{1}{m} \sum_{e \in C} S_e,
\]

where \( m \) is the number of elements in set \( C \), that is regarded as drawn randomly from a sample of \( E \) elements, either the total number of transcriptional profiled genes or the total number of semantic similarities across all the tested sets. Under the random-set scoring model (27) and thus conditional on element-level scores \( S_e \)

\[
\hat{\mu} = \frac{\sum_{e=1}^{E} S_e}{E}
\]

and

\[
\hat{\sigma}^2 = \frac{1}{m} \left( \frac{E - m}{E - 1} \right) \left\{ \left( \frac{\sum_{e=1}^{E} S_e^2}{E} \right) - \left( \frac{\sum_{e=1}^{E} S_e}{E} \right)^2 \right\}
\]

A standardized enrichment score (27) is then calculated as \( Z = (\tilde{X} - \hat{\mu})/\hat{\sigma} \), which under the null hypothesis “no association between the gene associated phenotypes and the disease phenotype” has expectation zero and unit variance (27). A large positive z-score favors the positive enrichment hypothesis that the candidate gene (or candidate complex) is strongly associated with mastitis. The candidate genes were ranked according to the standardized enrichment score.

**Phenome profiles of candidate complexes and candidate genes.** The difference between a gene-based and complex-based profile is that the biomedical records are linked to the candidate gene or to all the genes in the candidate complex. Thus the set of semantic similarities that is associated with a candidate gene is the gene-based phenotypic profile, while the set of semantic similarities associated with a candidate complex is the complex-based phenotypic profile. However, regardless of whether gene based or complex based, the phenotypic profile varies according to ortholog information and phenotype resources. Each cattle candidate gene was profiled according to 1) PubMed articles linked to the human ortholog, 2) OMIM records linked to the human ortholog, 3) GeneRIF records linked to the human ortholog, 4) GeneRIF linked to the mouse ortholog, and 5) GeneRIF records linked to the cattle gene. To this end, five categories of phenotype profiles are associated with each cattle candidate gene for gene based and complex based, respectively.

**Transcriptome profiles of candidate complexes and candidate genes.** For gene expression data, the cattle candidate complexes were based on human protein interactions that were derived from cattle candidate genes (see ortholog information and candidate complex). The transcriptional profile of the candidate complex made up by human orthologs was defined by the expression information (logFC) associated with the corresponding cattle genes. The gene-based profile was the t-statistics for each candidate gene according to the result of the expression analysis.

**Rank aggregation for combining phenome and integrating phenome with transcriptome.** The relative importance of the candidate genes for the disease was ranked according to the z-scores. As the gene-associated phenotypic profile can be defined by different sources of biomedical records (e.g., OMIM, PubMed, or GeneRIF) linked to different organisms (e.g., mouse or human) or from genome expression profiles, distinct phenotypic profiles will reveal the association between the disease and genes in a complementary dimension. Integration of the results was implemented by rank aggregation, which combines the ranking lists using different phenotypic profiles into a final rank. Rank aggregation (1) was used for combining the independent prioritization result that ranks genes according to disease relevance scores (z-scores) based on a specific category of gene-associated phenotype from text mining of biomedical text records or gene expression. The ranks of genes from these independent prioritizations, which represent the relative importance of genes to the disease, were transformed into rank ratios (divide the rank indexes by the number of candidate genes that were prioritized). For a given gene \( g \), its rank vector is defined as \( r(D,g) = [r_1, \ldots, r_{m}] \), where \( r_j \) corresponds to its rank ratio from an independent prioritization \( j \). The rank vector can be reordered increasingly to obtain the vector of order statistics \( r_{(1)} \leq \ldots \leq r_{(m)} \) which ranges from the smallest to the largest value of \( r(D,g) \). A small value indicates large relevance between disease and gene. The binomial distribution can be used to calculate the probability that \( k \) or more ranks are smaller than \( r_{(j)} \) for every \( k \) \( (1) \), which is the \( P \) value for each rank statistic \( r_{(j)} \) being randomly as small as observed:

\[
p_k = \sum_{j=k}^{m} \binom{m}{j} (r_{(j)})^j (1-r_{(j)})^{m-j}
\]

For each gene \( m \) probabilities was computed corresponding to one probability for each source of gene associated phenotype. Finally, the minimal probability (\( P \) value) was defined as the integrated disease relevance score of gene \( g \): \( p(D,g) = \min(p_1, \ldots, p_m) \)

This disease relevance score \( p \) computed for each gene was ordered increasingly to get the final prioritization result. Accordingly, a new ranking that aggregated individual prioritizations from the different gene associated phenotypic profiles from text mining of biomedical text records. This combined phenome result was then aggregated with the results according to gene-based and complex-based transcriptional profiles, respectively.

**Functional enrichment.** The functional enrichment analysis was performed for the top 100 genes in the complex-based transcriptome and the combined phenome using DAVID Bioinformatics Resources 6.7 (13, 14). The background for the selected gene sets was specified as “bovine array” for transcriptome and “Bos taurus” for phenome, respectively. The annotation terms [GO or Kyoto Encyclopedia of Genes and Genomes (KEGG)] were determined significantly based on two criteria. First, the modified Fisher exact \( P \) value was <0.01. Second, the threshold of minimum gene counts of an annotation term has to be \( \geq 2 \).

**RESULTS**

In this study we have implemented a network-based gene prioritization approach that integrates protein complexes, gene phenotype links, and gene-associated phenotypes obtained from text mining of biomedical records (e.g., OMIM, PubMed, and GeneRIF) linked to bovine genes or their orthologs in other species (e.g., human and mouse). We applied the network-
Based gene prioritization approach for ranking genes in relation to bovine mastitis by 1) global ranking of genes based on phenotype information, 2) global ranking of genes based on combined phenome and transcriptome information, and 3) local ranking of genes in a QTL region.

Global ranking of bovine genes based on gene-associated phenotypes obtained from text mining of biomedical text records. The network-based gene prioritization approach was used to compute a global ranking of all bovine genes according to their potential role in mastitis. Well-defined descriptions of phenotypes are fundamental for the accuracy in the prioritization of disease genes. Here, to rank bovine genes according to their potential role in mastitis, we used a disease phenotype description from Wikipedia as an analog to more authoritative peer-reviewed, reference-based resources (e.g., OMIM entry for human disorders). A phenotypic profile of individual genes was obtained from text mining of biomedical text records. Furthermore, phenome profiling requires that the candidate gene itself or its partners in the candidate complex are linked to one or more biomedical records. To capture the complementary information from various resources, we used the biomedical records from different sources, i.e., OMIM, PubMed, and GeneRIF, linked to genes from three different species (cattle, human, and mouse). Table 1 gives an overview of gene-associated phenotypic information in the various repositories. It is clear that ortholog mapping substantially increases the information available for gene prioritization (Fig. 2). Thus, by using gene-associated phenotype information from multiple species we were able to profile 57% of 22,486 bovine genes (Fig. 2). Compared with gene-based prioritizations there was only a small increase in the number of cattle genes that could be prioritized when information on candidate gene complexes were included (Fig. 2). However, we observed that the overlap of high-ranking genes between different sources of gene associated phenotypic profiles was larger when prioritizations were based on candidate complexes (Fig. 3). Thus, the number of unique genes in the union of all top 100 genes from different phenome profiles was 324 genes for the candidate complex-based prioritization compared with 412 genes in the candidate gene-based approach. The highest overlap (67%) was between prioritizations based on human gene complexes linked to PubMed abstracts and to GeneRIF records, whereas the lowest overlap (6%) was between prioritizations based on human gene complexes linked to GeneRIF and cattle gene complexes linked to GeneRIF. Prioritizations based on cattle genes linked to GeneRIF records seemed to be less consistent with those based on the other gene-associated phenotype profiles.

In Table 2, the highly prioritized genes (top 20) were sorted according to their rank aggregation by combining the individual prioritizations obtained using the various sources of biomedical records (see full list in Table S1 in Supplemental Data 1). Also shown are their ranks based on individual phenotype sources. The results clearly show that nearly all high-ranking genes are well-characterized immune genes such as members of IFNG signaling pathway, pathogen-associated molecular pattern (PAMP) receptors, and interleukins. Furthermore, a test for functional enrichment of GO Biological Process (BP) terms among the 100 highest ranking genes showed a strong enrichment (P value after false discovery rate correction is \( < 10^{-10} \)) of genes involved in inflammatory, immune, and defense responses (Table S2 in Supplemental Data 2). Also, functional enrichment analysis for KEGG pathways showed that the top 100 prioritized genes are significantly enriched in pathways such as cytokine-cytokine receptor interaction, the Jak-STAT signaling pathway, and the Toll-like receptor signaling pathway (Table 3). It is also noteworthy that 14 of the highly prioritized genes (Table 2) are found among the top 20 genes when ranking is based entirely on information in mouse GeneRIF.

Global ranking of bovine genes by combining information from gene-associated phenome profiles and transcriptome profiles. In the second application bovine genes were ranked based on combining information from phenotypic profiles obtained by text mining of biomedical text records and transcriptome profiles of udder tissue during E. coli infection (5).

---

Table 1. Overview of phenotype information provided by each resource with respect to species

<table>
<thead>
<tr>
<th>Phenotype Categories</th>
<th>Links, ( n )</th>
<th>Entrez Genes, ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human_PubMed</td>
<td>473,102</td>
<td>28,207</td>
</tr>
<tr>
<td>Human_OMIM</td>
<td>17,933</td>
<td>14,388</td>
</tr>
<tr>
<td>Human_GeneRIF</td>
<td>178,009</td>
<td>11,686</td>
</tr>
<tr>
<td>Mouse_GeneRIF</td>
<td>86,399</td>
<td>9,161</td>
</tr>
<tr>
<td>Cattle_GeneRIF</td>
<td>2,800</td>
<td>1,174</td>
</tr>
</tbody>
</table>

Five different resources supply the gene phenotype relationships independently. The statistics for the number of genes (human, mouse, or cattle) and number of links between a gene and a biomedical records (PubMed abstract, OMIM record, or GeneRIF record) are shown.
On a genome-wide scale, the transcriptional information for 11,800 genes and phenotypic information for 12,658 genes are available for prioritizing the relative disease importance accordingly. Based on that, the cattle genes were ordered by the rank aggregation of both types of information. The genes were ranked either as single genes (gene-based) according to the absolute level of the t-statistics or as a candidate complex (complex-based) according to z-scores computed from log-fold changes of the individual members in the complex. The transcriptome ranks was integrated with the ranks from the prioritizations based on phenome information. Thus, in Table 4 (and Table S3 in Supplemental Data 3), the genes were ordered by aggregation of the combined phenome (application 1) and the gene-based transcriptome rankings. The top 25 list contained 20 genes that could be ranked by both the phenome and the transcriptome and included seven genes (JAK2, IFNGR1, MYD88, SOCS3, TLR2, IL6, and IL1B) that ranked within top 20 in combined phenome. The top three genes (IL8, IL6R, and CCR1) were ranked within top 100 in both phenome and transcriptome. Interestingly, the top 20 genes also included three genes (IL8RB, KDM4B, and CSF3R), which ranked 1st, 2nd, and 4th in the transcriptome, but because of lack of biomedical records linked to these cattle genes and lack of their ortholog information in the annotation data packages they could not be prioritized in the combined phenome. Similarly, two genes (JAK2 and IFNGR1) were ranked 1st and 2nd in phenome, but they could not be ranked based on transcriptome information due to the absence of corresponding probe-sets on the Bovine Genome array.

Table 5 shows the top 20 genes prioritized by aggregation of the combined phenome and complex-based transcriptome rankings (see full list in Table S4 in Supplemental Data 4). The list contains 10 genes (IL1B, NOS2, IL6, TLR4, IFNG, IL1A, IL5, IL18, IL22, and IL10), which were highly ranked in top 20 in combined phenome. These 20 genes were also found among the top 150 highest ranking genes in both phenome and transcriptome. Four genes (IL1B, IL6, IL6R, and ICAM1) were highly ranked according to aggregation of phenome and gene-based transcriptome rankings (Table 4) as well as according to aggregation of the phenome and complex-based transcriptome rankings (Table 5).

Table 2. Top 20 genes in the rank list of combined phenome

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>JAK2</td>
<td>1.83E-09</td>
<td>66</td>
<td>6</td>
<td>137</td>
<td>28</td>
<td>7</td>
<td>Janus kinase 2</td>
</tr>
<tr>
<td>2</td>
<td>IFNGR1</td>
<td>2.11E-09</td>
<td>79</td>
<td>14</td>
<td>141</td>
<td>10</td>
<td>16</td>
<td>interferon gamma receptor 1</td>
</tr>
<tr>
<td>3</td>
<td>MYD88</td>
<td>8.33E-09</td>
<td>31</td>
<td>21</td>
<td>49</td>
<td>7</td>
<td>385</td>
<td>myeloid differentiation primary response gene (88)</td>
</tr>
<tr>
<td>4</td>
<td>TLR2</td>
<td>8.45E-09</td>
<td>19</td>
<td>57</td>
<td>41</td>
<td>8</td>
<td>34</td>
<td>toll-like receptor 2</td>
</tr>
<tr>
<td>5</td>
<td>IL6</td>
<td>1.06E-08</td>
<td>22</td>
<td>16</td>
<td>52</td>
<td>5</td>
<td>188</td>
<td>interleukin 6 (interferon, beta 2)</td>
</tr>
<tr>
<td>6</td>
<td>IL10</td>
<td>1.23E-08</td>
<td>15</td>
<td>17</td>
<td>54</td>
<td>11</td>
<td>31</td>
<td>interleukin 10</td>
</tr>
<tr>
<td>7</td>
<td>IL1B</td>
<td>1.32E-08</td>
<td>18</td>
<td>28</td>
<td>55</td>
<td>3</td>
<td>189</td>
<td>interleukin 1, beta</td>
</tr>
<tr>
<td>8</td>
<td>SOCS3</td>
<td>1.52E-08</td>
<td>50</td>
<td>4</td>
<td>57</td>
<td>9</td>
<td>58</td>
<td>suppressor of cytokine signaling 3</td>
</tr>
<tr>
<td>9</td>
<td>IL1A</td>
<td>1.63E-08</td>
<td>27</td>
<td>45</td>
<td>58</td>
<td>31</td>
<td>943</td>
<td>interleukin 1, alpha</td>
</tr>
<tr>
<td>10</td>
<td>JIF</td>
<td>4.37E-08</td>
<td>94</td>
<td>86</td>
<td>66</td>
<td>46</td>
<td>439</td>
<td>leukemia inhibitory factor (cholinergic differentiation factor)</td>
</tr>
<tr>
<td>11</td>
<td>IL18</td>
<td>5.90E-08</td>
<td>26</td>
<td>84</td>
<td>80</td>
<td>13</td>
<td>830</td>
<td>interleukin 18 (interferon-gamma-inducing factor)</td>
</tr>
<tr>
<td>12</td>
<td>IL13</td>
<td>6.77E-08</td>
<td>21</td>
<td>96</td>
<td>68</td>
<td>6</td>
<td>305</td>
<td>interleukin 13</td>
</tr>
<tr>
<td>13</td>
<td>IL23A</td>
<td>7.17E-08</td>
<td>29</td>
<td>26</td>
<td>84</td>
<td>19</td>
<td>731</td>
<td>interleukin 23, alpha subunit p19</td>
</tr>
<tr>
<td>14</td>
<td>IL5</td>
<td>8.25E-08</td>
<td>25</td>
<td>70</td>
<td>87</td>
<td>15</td>
<td>NA*</td>
<td>interleukin 5 (colony-stimulating factor, cosinophil)</td>
</tr>
<tr>
<td>15</td>
<td>IFNG</td>
<td>8.63E-08</td>
<td>44</td>
<td>7</td>
<td>88</td>
<td>2</td>
<td>86</td>
<td>interferon, gamma</td>
</tr>
<tr>
<td>16</td>
<td>TLR4</td>
<td>1.12E-07</td>
<td>36</td>
<td>41</td>
<td>94</td>
<td>16</td>
<td>141</td>
<td>toll-like receptor 4</td>
</tr>
<tr>
<td>17</td>
<td>PALB2</td>
<td>1.70E-07</td>
<td>1</td>
<td>87</td>
<td>1</td>
<td>NA*</td>
<td>NA*</td>
<td>partner and localizer of BRCA2</td>
</tr>
<tr>
<td>18</td>
<td>NOS2</td>
<td>1.82E-07</td>
<td>34</td>
<td>32</td>
<td>100</td>
<td>83</td>
<td>985</td>
<td>nitric oxide synthase 2, inducible</td>
</tr>
<tr>
<td>19</td>
<td>IL17A</td>
<td>1.97E-07</td>
<td>28</td>
<td>24</td>
<td>128</td>
<td>1</td>
<td>NA*</td>
<td>interleukin 17A</td>
</tr>
<tr>
<td>20</td>
<td>IL22</td>
<td>2.18E-07</td>
<td>89</td>
<td>69</td>
<td>111</td>
<td>45</td>
<td>NA*</td>
<td>interleukin 22</td>
</tr>
</tbody>
</table>

Shown are the top 20 genes based on the combined results given by 5 different phenotype sources of complex based. Genes are ordered according to the minimal P values derived from rank aggregation of human PubMed, OMIM, and GeneRIF of human (Hsa), mouse (Mmu), and cattle (Bta). The individual ranks using distinct phenotype sources are given correspondingly. *Gene can not be ranked in the corresponding category.
The relationship between prioritizations using different sources of gene-associated phenotype information from text mining of biomedical text records and transcriptome profiling was investigated (Fig. 4). The top 100 genes were selected either based on combined phenotype rankings (left) or complex-based transcriptome rankings (right). The diagonal elements of the triangular matrix show the numbers of top prioritized genes for each profile and the off-diagonals the number of overlapping genes among the different profiles.

The left triangular matrix in Fig. 4 shows that among the top 100 genes ranked according to the combined phenotype, a set of 49 genes (~50%) were confirmed to be associated to bovine mastitis by our independent gene expression study (5). Nearly all of these 100 genes were predicted to be highly associated with bovine mastitis using different sources of biomedical records including human PubMed (100%), OMIM (96%), or human GeneRIF (96%). Although fewer (77%) and even fewer (14%) of these genes were prioritized as the top candidates using mouse and cattle GeneRIF, respectively, the proportions in terms of common set with the genes supported by transcriptome are interestingly higher in mouse 57% (44/77) and cattle 71% (10/14), compared with that using human phenotypic information (~50%).

The high-ranking genes from the transcriptome profiling (right triangular matrix in Fig. 4) is supported by the gene-associated phenotypes obtained from text mining of biomedical text records, except for cattle GeneRIF in which no genes were prioritized to be significantly associated with mastitis. For

Table 3. **KEGG enrichment of top 100 genes from combined phenome**

<table>
<thead>
<tr>
<th>KEGG_ID</th>
<th>KEGG_Name</th>
<th>Count</th>
<th>P Value</th>
<th>Bonferroni</th>
<th>Benjamini</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>bta04060</td>
<td>Cytokine-cytokine receptor interaction</td>
<td>38</td>
<td>9.93E-34</td>
<td>6.55E-32</td>
<td>6.55E-32</td>
<td>1.02E-30</td>
</tr>
<tr>
<td>bta04620</td>
<td>Toll-like receptor signaling pathway</td>
<td>17</td>
<td>3.43E-13</td>
<td>2.26E-11</td>
<td>2.26E-11</td>
<td>3.50E-10</td>
</tr>
<tr>
<td>bta04672</td>
<td>Intestinal immune network for IgA production</td>
<td>11</td>
<td>1.33E-09</td>
<td>8.77E-08</td>
<td>8.77E-08</td>
<td>1.36E-06</td>
</tr>
<tr>
<td>bta05332</td>
<td>Graft-versus-host disease</td>
<td>9</td>
<td>9.62E-09</td>
<td>6.35E-07</td>
<td>6.35E-07</td>
<td>9.84E-06</td>
</tr>
<tr>
<td>bta05330</td>
<td>Allograft rejection</td>
<td>9</td>
<td>2.84E-08</td>
<td>1.87E-06</td>
<td>1.87E-06</td>
<td>2.90E-05</td>
</tr>
<tr>
<td>bta04621</td>
<td>NOD-like receptor signaling pathway</td>
<td>10</td>
<td>7.66E-08</td>
<td>5.06E-06</td>
<td>5.06E-06</td>
<td>7.83E-05</td>
</tr>
<tr>
<td>bta04940</td>
<td>type 1 diabetes mellitus</td>
<td>9</td>
<td>9.05E-08</td>
<td>5.97E-06</td>
<td>5.97E-06</td>
<td>9.25E-05</td>
</tr>
<tr>
<td>bta04940</td>
<td>Hematopoietic cell lineage</td>
<td>10</td>
<td>2.32E-06</td>
<td>0.00153</td>
<td>0.00153</td>
<td>0.002368</td>
</tr>
<tr>
<td>bta05320</td>
<td>Autoimmune thyroid disease</td>
<td>8</td>
<td>3.19E-06</td>
<td>0.00021</td>
<td>0.00021</td>
<td>0.003268</td>
</tr>
<tr>
<td>bta04710</td>
<td>Apoptosis</td>
<td>9</td>
<td>3.08E-05</td>
<td>0.002033</td>
<td>0.002033</td>
<td>0.003151</td>
</tr>
<tr>
<td>bta04920</td>
<td>Adipocytokine signaling pathway</td>
<td>8</td>
<td>4.28E-05</td>
<td>0.002322</td>
<td>0.002322</td>
<td>0.004372</td>
</tr>
<tr>
<td>bta05660</td>
<td>T cell receptor signaling pathway</td>
<td>9</td>
<td>0.000182</td>
<td>0.000922</td>
<td>0.000922</td>
<td>0.015522</td>
</tr>
<tr>
<td>bta05310</td>
<td>Asthma</td>
<td>5</td>
<td>0.000435</td>
<td>0.002822</td>
<td>0.002822</td>
<td>0.004417</td>
</tr>
<tr>
<td>bta04623</td>
<td>Cytosolic DNA-sensing pathway</td>
<td>6</td>
<td>0.000849</td>
<td>0.005401</td>
<td>0.005401</td>
<td>0.008442</td>
</tr>
</tbody>
</table>

Shown are the KEGG pathways enriched (P < 0.01) in top 100 genes of combined phenome. The number (count) of genes comprising each KEGG pathway and adjusted results using different methods for multiple test correction are also given. FDR, false discovery rate.

Table 4. **Top 25 genes in the rank list integrating combined phenome and gene-based transcriptome**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Symbol</th>
<th>Min. P Value</th>
<th>Transcriptome</th>
<th>Phenome</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IL8</td>
<td>3.33E-05</td>
<td>30</td>
<td>73</td>
<td>Interleukin 8</td>
</tr>
<tr>
<td>2</td>
<td>IL6R</td>
<td>3.99E-05</td>
<td>72</td>
<td>79</td>
<td>Interleukin 6 receptor</td>
</tr>
<tr>
<td>3</td>
<td>CCR1</td>
<td>5.87E-05</td>
<td>10</td>
<td>97</td>
<td>Chemokine (C-C motif) receptor 1</td>
</tr>
<tr>
<td>4</td>
<td>ICAM1</td>
<td>0.000111</td>
<td>120</td>
<td>45</td>
<td>Interleukin 5, receptor 1</td>
</tr>
<tr>
<td>5</td>
<td>JAK2</td>
<td>0.000158</td>
<td>NA</td>
<td>1</td>
<td>Janus kinase 2</td>
</tr>
<tr>
<td>6</td>
<td>IL8RB</td>
<td>0.000175</td>
<td>1</td>
<td>NA</td>
<td>Interleukin 8 receptor, beta</td>
</tr>
<tr>
<td>7</td>
<td>IL1R2</td>
<td>0.000196</td>
<td>6</td>
<td>177</td>
<td>Interleukin 1 receptor, type II</td>
</tr>
<tr>
<td>8</td>
<td>NFKB1A</td>
<td>0.000255</td>
<td>114</td>
<td>202</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha</td>
</tr>
<tr>
<td>9</td>
<td>IFNGR1</td>
<td>0.000316</td>
<td>NA</td>
<td>2</td>
<td>Interferon gamma receptor 1</td>
</tr>
<tr>
<td>10</td>
<td>KDM4B</td>
<td>0.000351</td>
<td>2</td>
<td>NA</td>
<td>Lysine (K)-specific demethylase 4B</td>
</tr>
<tr>
<td>11</td>
<td>SOCS2</td>
<td>0.000452</td>
<td>128</td>
<td>269</td>
<td>Suppressor of cytokine signaling 2</td>
</tr>
<tr>
<td>12</td>
<td>PYCARD</td>
<td>0.000458</td>
<td>244</td>
<td>72</td>
<td>PYD and CARD domain containing</td>
</tr>
<tr>
<td>13</td>
<td>MYD88</td>
<td>0.000474</td>
<td>723</td>
<td>5</td>
<td>Myeloid differentiation primary response gene (88)</td>
</tr>
<tr>
<td>14</td>
<td>TRA1</td>
<td>0.000508</td>
<td>257</td>
<td>110</td>
<td>Tumor rejection antigen (gp96) 1</td>
</tr>
<tr>
<td>15</td>
<td>SOCS3</td>
<td>0.000525</td>
<td>260</td>
<td>8</td>
<td>Suppressor of cytokine signaling 3</td>
</tr>
<tr>
<td>16</td>
<td>MGC165939</td>
<td>0.000526</td>
<td>3</td>
<td>11734</td>
<td>Uncharacterized protein C13orf18 homolog</td>
</tr>
<tr>
<td>17</td>
<td>TLR2</td>
<td>0.000632</td>
<td>390</td>
<td>4</td>
<td>Toll-like receptor 2</td>
</tr>
<tr>
<td>18</td>
<td>CCL19</td>
<td>0.000675</td>
<td>68</td>
<td>329</td>
<td>Chemokine (C-C motif) ligand 19</td>
</tr>
<tr>
<td>19</td>
<td>CSF3R</td>
<td>0.000702</td>
<td>4</td>
<td>NA</td>
<td>Similar to colony stimulating factor 3 receptor</td>
</tr>
<tr>
<td>20</td>
<td>CXCL16</td>
<td>0.000732</td>
<td>302</td>
<td>342</td>
<td>Chemokine (C-X-C motif) ligand 16</td>
</tr>
<tr>
<td>21</td>
<td>IL6</td>
<td>0.000759</td>
<td>793</td>
<td>5</td>
<td>Interleukin 6 (interferon, beta 2)</td>
</tr>
<tr>
<td>22</td>
<td>IL4R</td>
<td>0.000798</td>
<td>322</td>
<td>82</td>
<td>Interleukin 4 receptor</td>
</tr>
<tr>
<td>23</td>
<td>CD40</td>
<td>0.000805</td>
<td>119</td>
<td>385</td>
<td>CD40 molecule, TNF receptor superfamily member 5</td>
</tr>
<tr>
<td>24</td>
<td>IL1B</td>
<td>0.000869</td>
<td>336</td>
<td>7</td>
<td>Interleukin 1, beta</td>
</tr>
<tr>
<td>25</td>
<td>IL2RA</td>
<td>0.000873</td>
<td>85</td>
<td>374</td>
<td>Interleukin 2 receptor, alpha</td>
</tr>
</tbody>
</table>

Listed are the top 25 genes based on the integrated results given by the combined phenome and gene-based transcriptome. Genes are ordered according to the minimal P values derived from rank aggregation of combined phenome and gene-based transcriptome, of which the individual ranks are given correspondingly.

*Gene can not be ranked in the corresponding category.*
example, 24 and 21 genes were prioritized as the top candidates according to human PubMed and OMIM, respectively, and a common set of 16 genes was observed by both. Twelve genes (IL1B, NOS2, IL6, TLR4, IFNG, MMP9, TNFSF11, TYK2, IL1A, SELE, IL6R, and IL5) were supported by all phenotype sources that implemented ortholog mapping (human PubMed, human OMIM, human GeneRIF, and mouse GeneRIF). It is notable that 69% of the top differentially expressed genes were not strongly associated with mastitis based on the prioritizations using current phenotype information.

The majority of the BP (70%) and KEGG pathways (60%) that were enriched among the top 100 genes of transcriptome (Table 6 and Table 7) were also enriched among the top 100 genes ranked according to the combined phenome information (Table S2 in Supplemental Data 2 and Table 3). KEGG pathways involved in “glycerophospholipid metabolism” and “leukocyte transendothelial migration” could only be revealed based on the transcriptional profile. Biological processes that were only identified from the transcriptome information include “response to fungus,” “chemotaxis,” “taxis,” “alcohol biosynthetic process,” “defense response to fungus,” “oxygen and reactive oxygen species metabolic process,” “cellular carbohydrate biosynthetic process,” “superoxide metabolic process,” “locomotory behavior,” “carbohydrate biosynthetic process,” “gluconeogenesis,” “behavior,” and “pyruvate metabolic process.” More than half of these biological processes enriched in the expression study are related to metabolic activities.

### Table 5. Top 20 genes in the rank list of integrating combined phenome and complex-based transcriptome

<table>
<thead>
<tr>
<th>Rank</th>
<th>Symbol</th>
<th>Min. P Value</th>
<th>Transcriptome</th>
<th>Phenome</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IL1B</td>
<td>4.14E-06</td>
<td>24</td>
<td>7</td>
<td>interleukin 1, beta</td>
</tr>
<tr>
<td>2</td>
<td>NOS2</td>
<td>1.27E-05</td>
<td>42</td>
<td>18</td>
<td>nitric oxide synthase 2, inducible</td>
</tr>
<tr>
<td>3</td>
<td>IL6</td>
<td>2.02E-05</td>
<td>53</td>
<td>5</td>
<td>interferon, beta 2</td>
</tr>
<tr>
<td>4</td>
<td>TLR4</td>
<td>2.17E-05</td>
<td>55</td>
<td>16</td>
<td>toll-like receptor 4</td>
</tr>
<tr>
<td>5</td>
<td>IFNG</td>
<td>2.33E-05</td>
<td>57</td>
<td>15</td>
<td>interferon, gamma</td>
</tr>
<tr>
<td>6</td>
<td>TYK2</td>
<td>3.93E-05</td>
<td>74</td>
<td>78</td>
<td>tyrosine kinase 2</td>
</tr>
<tr>
<td>7</td>
<td>IL1A</td>
<td>4.71E-05</td>
<td>81</td>
<td>9</td>
<td>interleukin 1, alpha</td>
</tr>
<tr>
<td>8</td>
<td>SELE</td>
<td>5.19E-05</td>
<td>85</td>
<td>81</td>
<td>selectin E</td>
</tr>
<tr>
<td>9</td>
<td>IL6R</td>
<td>5.56E-05</td>
<td>88</td>
<td>79</td>
<td>interleukin 6 receptor</td>
</tr>
<tr>
<td>10</td>
<td>TNFSF11</td>
<td>5.75E-05</td>
<td>70</td>
<td>96</td>
<td>tumor necrosis factor (ligand) superfamily, member 11</td>
</tr>
<tr>
<td>11</td>
<td>IL5</td>
<td>6.48E-05</td>
<td>95</td>
<td>14</td>
<td>interleukin 5 (colony-stimulating factor, eosinophil)</td>
</tr>
<tr>
<td>12</td>
<td>MMP9</td>
<td>7.41E-05</td>
<td>59</td>
<td>109</td>
<td>matrix metalloproteinase 9 (gelatinase B, 92 kDa gelatinase, 92 kDa type IV collagenase)</td>
</tr>
<tr>
<td>13</td>
<td>IL18</td>
<td>8.85E-05</td>
<td>111</td>
<td>11</td>
<td>interleukin 18 (interferon-gamma-inducing factor)</td>
</tr>
<tr>
<td>14</td>
<td>IL22</td>
<td>9.66E-05</td>
<td>116</td>
<td>20</td>
<td>interleukin 22</td>
</tr>
<tr>
<td>15</td>
<td>IL10</td>
<td>0.000116</td>
<td>127</td>
<td>6</td>
<td>interleukin 10</td>
</tr>
<tr>
<td>16</td>
<td>ICAM1</td>
<td>0.000119</td>
<td>129</td>
<td>45</td>
<td>intercellular adhesion molecule 1</td>
</tr>
<tr>
<td>17</td>
<td>CXCL12</td>
<td>0.000123</td>
<td>131</td>
<td>111</td>
<td>chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)</td>
</tr>
<tr>
<td>18</td>
<td>JAK1</td>
<td>0.000125</td>
<td>132</td>
<td>86</td>
<td>Janus kinase 1</td>
</tr>
<tr>
<td>19</td>
<td>CD44</td>
<td>0.000135</td>
<td>137</td>
<td>41</td>
<td>CD44 molecule (Indian blood group)</td>
</tr>
<tr>
<td>20</td>
<td>DCN</td>
<td>0.00014</td>
<td>126</td>
<td>150</td>
<td>decorin</td>
</tr>
</tbody>
</table>

Listed are the top 20 genes based on the integrated results given by the combined phenome and complex-based transcriptome. Genes are ordered according to the minimal P values derived from rank aggregation of combined phenome and complex-based transcriptome, of which the individual ranks are given correspondingly.
KEGG pathway and adjusted results using different methods for multiple test correction are also given.

<table>
<thead>
<tr>
<th>KEGG_ID</th>
<th>KEGG_Name</th>
<th>Count</th>
<th>P Value</th>
<th>Bonferroni</th>
<th>Benjamini</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>bta04640</td>
<td>hematopoietic cell lineage</td>
<td>6</td>
<td>0.000477</td>
<td>0.025424</td>
<td>0.025424</td>
<td>0.46612</td>
</tr>
<tr>
<td>bta04600</td>
<td>cytokine-cytokine receptor interaction</td>
<td>7</td>
<td>0.001688</td>
<td>0.087195</td>
<td>0.044592</td>
<td>1.641509</td>
</tr>
<tr>
<td>bta04670</td>
<td>leukocyte transendothelial migration</td>
<td>6</td>
<td>0.002302</td>
<td>0.117032</td>
<td>0.005654</td>
<td>0.095775</td>
</tr>
<tr>
<td>bta05332</td>
<td>graft-versus-host disease</td>
<td>4</td>
<td>0.003767</td>
<td>0.184379</td>
<td>0.049675</td>
<td>3.629865</td>
</tr>
<tr>
<td>bta05654</td>
<td>glycerophospholipid metabolism</td>
<td>4</td>
<td>0.009754</td>
<td>0.410982</td>
<td>0.100449</td>
<td>9.155782</td>
</tr>
</tbody>
</table>

Listed are the KEGG pathways enriched (P < 0.01) in the top 100 genes of complex-based transcriptome. The number (count) of genes comprised in each KEGG pathway and adjusted results using different methods for multiple test correction are also given.

Table 8 (full list in Table S5 in Supplemental Data 5). *IFNGR1* ranked highest among the genes located in the QTL region. On a genome-wide scale *IFNGR1* also appeared to be implicated in mastitis since it ranked second based on phenome information alone and 39th when individual ranks from the phenome and transcriptome prioritizations were combined. The 2nd and 3rd highest ranking genes in the QTL region were VNN1 and VNN2, which also on a global scale ranked among the 160 most mastitis relevant (~top 1%) genes.

**DISCUSSION**

Network-based gene prioritization approach. In this study we have developed a network-based gene prioritization ap-
proach for ranking genes associated with diseases in livestock species. Our prioritization approach uses ortholog mapping and integrates information from disease phenotypes, gene-associated phenotypes, and PPI in the prioritizations. It provides a general framework for identifying and ranking genes associated with complex diseases in livestock populations, helping efforts to improve animal health and welfare. It can be used to prioritize candidate genes, for example genes that are differentially expressed, located in a specific genomic region, or the global set of genes present in the genome. Thus, our network-based gene prioritization approach complements high-throughput experimental approaches such as genome-wide association analyses and genome-wide expression analyses.

In the present study, the approach was used to identify and rank bovine genes according to their potential role in bovine mastitis. Based on phenome prioritizations, the high-ranking candidates include pathogen-associated molecular pattern (PAMP) receptors (e.g., TLR2 and TLR4) as well as pro- and anti-inflammatory cytokines (e.g., IL1A, IL1B, IL6, and IL10), which are genes well known to be involved in inflammatory response directly toward bacteria in bovine mastitis (31, 32). The high-ranking genes also consist of a number of interleukins, IFNG, IFNGR1, NOS2, PALB2, and participants in signaling cascade (e.g., JAK2, MYD88, and SOCS3) and acute-phase response (e.g., LIF). Combining information from phenotype and transcriptome profiles also revealed some novel discoveries (see genes in Tables S3 and S4 in Supplemental Data). Local prioritization of genes in a QTL region previous identified as associated with mastitis risk in several dairy breeds pointed to IFNGR1 as the most likely gene to harbor genetic variants affecting susceptibility/resistance to mastitis. The inflammatory response during mastitis can be different when induced by gram-positive bacteria (e.g., S. aureus) compared with that induced by gram-negative bacteria (e.g., E. coli) (3). Regarding IFNG, it has been shown to be upregulated in response to both classes of pathogens in acute and chronic phases when investigated in experimental mastitis models. In the present study IFNG was ranked 15th based on text mining of biomedical records and 57th based on the expression of profile of its interaction partners in the candidate complex. The IFNGR1 gene, which was ranked second according to the phenotype profiles, encodes the ligand-binding chain (alpha) of the IFNG receptor. However, the probe-set for detecting IFNGR1 is not present on the Bovine Genome arrays, thus the relative expression level of IFNGR1 could not be determined in expression study. In humans, polymorphisms in the IFNGR1 gene are associated with susceptibility to Helicobacter pylori infection (40). Also, mutations in the IFNGR1 gene have been shown to be caused by high susceptibility to mycobacterial disease, underscoring the critical role of the interferon pathway in response to pathogens (19). The genes VNN1 and VNN2 (panetheinase gene family) were also found to be functionally related to resistance to bovine mastitis. Mice lacking VNN1 are protected from acute inflammation triggered by toxic or infectious insults in gut. In contrast these mice poorly develop organized granulomas in response to infection by intracellular bacteria. This suggests that VNN1 expressed by epithelial cells exerts a control on immune responses (4, 24).

**Data integration.** We used a rank order aggregation method for integration of information from heterogeneous data sources. This method seems suitable for combining different prioritizations that are distinct in many aspects, for instance the level of data and biological meaningfulness. Future developments may include alternative integrating approaches or additional weighting schemes that consider the quantity and quality of the individual prioritization as well as their relationship.

Here, we integrated transcriptome, phenome, and interactome data to rank bovine genes according to their potential role in bovine mastitis. The prioritizations that were based on candidate gene complexes appeared to be more consistent between different phenotype sources and also between transcriptome and phenome than the prioritizations based on candidate genes only. This suggests that the PPI play an important role in data integration, most likely because it takes into account the contributions of distinct genes to the trait of interest.

The transcriptome data provide a large number of intermediate phenotypes that are heavily influenced by the origin of tissue, chronology in the pathogenic process, and posttranscriptional modification (mRNA processing, splicing, and microRNA-mediated regulation, etc.) in the cells. These intermediate phenotypes may influence the ultimate phenotypic characteristics of the disease traits. Functional enrichment tests showed an overrepresentation of similar immunological processes among the genes that ranked high in both the phenome and the transcriptome. In contrast, genes involved in metabolism were

### Table 8. Top 10 genes in QTL regions ranked according to combined phenome and complex-based transcriptome

<table>
<thead>
<tr>
<th>Rank</th>
<th>Symbol</th>
<th>Integrated Rank</th>
<th>Transcriptome Rank</th>
<th>Phenome Rank</th>
<th>Entrezgene</th>
<th>Name</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IFNGR1</td>
<td>39</td>
<td>860</td>
<td>2</td>
<td>508619</td>
<td>interferon gamma receptor 1</td>
<td>7738543</td>
<td>77409861</td>
</tr>
<tr>
<td>2</td>
<td>VNN1</td>
<td>143</td>
<td>178</td>
<td>636</td>
<td>526704</td>
<td>vanin 1</td>
<td>72929959</td>
<td>72948513</td>
</tr>
<tr>
<td>3</td>
<td>VNN2</td>
<td>160</td>
<td>18</td>
<td>4831</td>
<td>534929</td>
<td>vanin 2</td>
<td>7291240</td>
<td>73011194</td>
</tr>
<tr>
<td>4</td>
<td>IL22RA2</td>
<td>218</td>
<td>867</td>
<td>561</td>
<td>533061</td>
<td>interleukin 22 receptor, alpha 2</td>
<td>77357952</td>
<td>77377309</td>
</tr>
<tr>
<td>5</td>
<td>MAP3K7</td>
<td>667</td>
<td>1852</td>
<td>708</td>
<td>529146</td>
<td>mitogen-activated protein kinase kinase 7 62341144</td>
<td>62404060</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>IL20RA</td>
<td>679</td>
<td>1865</td>
<td>408</td>
<td>509038</td>
<td>interleukin 20 receptor, alpha</td>
<td>77272571</td>
<td>77350927</td>
</tr>
<tr>
<td>7</td>
<td>TNFAIP3</td>
<td>888</td>
<td>2209</td>
<td>1892</td>
<td>508105</td>
<td>tumor necrosis factor, alpha-induced protein 3</td>
<td>78050571</td>
<td>78060888</td>
</tr>
<tr>
<td>8</td>
<td>CGA</td>
<td>890</td>
<td>11401</td>
<td>224</td>
<td>280749</td>
<td>glycoprotein hormones, alpha polypeptide</td>
<td>64631036</td>
<td>64647692</td>
</tr>
<tr>
<td>9</td>
<td>ENPP3</td>
<td>1154</td>
<td>296</td>
<td>NA</td>
<td>529405</td>
<td>ectonucleotide pyrophosphatase/phosphodiesterase 3</td>
<td>71707032</td>
<td>71795137</td>
</tr>
<tr>
<td>10</td>
<td>HEBP2</td>
<td>1296</td>
<td>1779</td>
<td>3067</td>
<td>509223</td>
<td>heme binding protein 2</td>
<td>78495672</td>
<td>78502467</td>
</tr>
</tbody>
</table>
only overrepresented when rankings was based transcriptome information, thus illustrating a cross-link between metabolic activity and immunological processes that is not captured by text mining of biomedical records. This may be partially explained by the dependency of phenome prioritization on the phenotype information of both genes and the target disease. If the text description of disease phenotype does not contain specific information on metabolism, the relevant biomedical records will not be retrieved and the genes related to metabolism will not be highly ranked. On the other hand the description of the disease phenotype can also be used as a filter to select genes from candidate list that possess predefined trait features such as immunological activity. Consequently, it can help to distinguish between differentially expressed genes that may be directly affected by mastitis infection (genes engaged in inflammatory response) from genes that were indirectly affected (genes associated with metabolic activity). Thus we believe that gene prioritization based on the text mining of biomedical text records provides a systematic and quantitative way of ranking candidate genes, which is superior to manually searching for disease relevance of genes in the scientific literature. Also, it may enable novel annotation of genes by associating genes with specific phenotypes or traits. Overall the integration of information from both transcriptome and phenotype is a tradeoff between discovering specificity and generality of a gene’s contribution to disease, which we believe eventually will help us to a better understanding of complex traits.

**Disease phenotypes.** The prioritizations of disease genes rely on the accuracy of the phenotypic description of the trait or disease. Ideally, the textual description of the disease phenotype should be based on expert reviewing of information retrieved from the published literature such as those of the OMIM database for human diseases. However, the content of the analog database in animals, OMIA (28), is still very limited and mastitis is not represented. Therefore, we considered alternative sources of information on disease phenotypes such as Wikipedia or Merck Veterinary Manual (http://www.merckvetmanual.com/mvm/index.jsp). To demonstrate the gene prioritization approach and to ensure the validity and reproducibility, we took the wiki page of mastitis as an example. Wiki enables communities to write documents collaboratively and provide means and control for verification, which is similar to processes involved in creating the expert-curated disease descriptions in the OMIM database. Indeed, our results show that the phenotypic description provided by the Wikipedia article exhibits a remarkable accuracy because the high-ranking genes are highly enriched for biological processes associated with inflammatory, immune, and defense responses. To further improve the specificity and sensitivity for detecting and ranking genes relevant for animal diseases, future steps may include the use of vocabulary filters that relate specifically to veterinary medicine, such as veterinary terminology. In addition, a customized trait description for livestock compiled by veterinarians/biologists, or alternatively a collection of full-text publications on the disease of interest, which may also enhance the accuracy of genes prioritization for animal diseases.

**Phenotypic profile defined by different sources of gene-associated phenotypes.** To use our network-based gene prioritization approach for ranking candidate genes it is necessary that the gene itself or its partners in the complex are linked to one or more biomedical records. We showed that there was a large common set (up to 67%) in the disease gene prioritization using different sources of gene-associated phenotypes (OMIM, PubMed, and GeneRIF). Despite the consistency, it probably reflects the functional annotations in GeneRIF being linked to genes described in PubMed. On the other site, the findings of less agreement between different prioritizations suggest that phenotypic profiling using various biomedical records provides complementary information in the sense that it allows us to comprehensively capture gene phenotype relationships from different biomedical databases. Furthermore, it allows us to collect phenotypic evidence from different species including human and mouse. This is important since the mouse has been widely used as an experimental model in the field of biomedicine and thus provides many links between mouse genes and a wide variety of trait phenotypes. A large amount of experimental trials (e.g., host responses to pathogens) have been performed in mice, leading to new discoveries and gene-phenotype associations, which are published in biomedical repositories. The impact of phenotypic evidence collected in mouse was demonstrated in our global ranking results (Table 2), where we found that 14 out of the 20 highest ranking genes were strongly supported by the rankings based solely on mouse information. Prioritizations based on cattle genes linked to GeneRIF records seemed to be less consistent with prioritizations based on the other phenotype sources. This is mainly attributed to the limited number of biomedical records linked to cattle genes in the current GeneRIF database. However, as the number of genome-wide studies of diseases and quantitative traits in dairy cattle are rapidly growing we expect that the updated literature source (GeneRIF or PubMed) eventually will provide more complete phenotypic information and a better coverage on cattle genes.

**Ortholog information.** An important feature of our network-based gene prioritization approach is that it allow us to combine information from multiple species (cattle, human, and mouse). The ortholog mapping is an essential step as it results in approximately a 10-fold increase in the number of cattle genes that can be ranked. In addition, the phenotypic information available for ranking is much higher in humans and mouse compared with cattle since the average number of links between a gene and GeneRIF records is four to six times higher. Ortholog mapping also increases the number of PPI, enabling us to capture more biomedical records in the analyses. This may in turn explain that the prioritization was more consistent in the complex-based approach compared with the gene-based approach. This is in line with complex-based disease gene prioritizations approaches used in human studies (7, 9, 10, 18, 20, 42–45). Compared with human and model organisms (e.g., mouse, yeast, and worm), much fewer protein–protein interactions and gene-associated phenotypes are available in cattle. Since the molecular and biochemical functions of a given protein are generally conserved between organisms, one can infer PPI and phenotype information about cattle genes by mapping to the ortholog genes in human or mouse. The prioritizations are based on transfer of phenotypic information between species, which contribute substantially to prioritizing cattle genes. However, it should be noted that the functional disruption of a conserved gene may give rise to radically different phenotypic outcomes in different species (25). That
is, although a gene’s “molecular” functions are conserved, the “organism-level” functions need not be (25). When a conserved gene is mutated, the ultimate phenotypes at organism level is an emergent property of the system (25). This concern will not entirely negate the effort of ortholog mapping to infer more phenotypic information for cattle genes, but it does provide a note of caution. However, we implemented phenotypic profiling in a network-based context, which is likely to be more robust as the phenotypes were represented by a functional module rather than a single gene. The orthogonal gene-associated phenotypes used in this study come from different biomedical databases (e.g., OMIM records, PubMed abstracts, and GeneRIF records) of different levels on several species including those linked directly to cattle genes. Hence, the interspecies difference in phenotype is unlikely to cause a large decrease of confidence in gene prioritizations in the present study.

Conclusion

In conclusion, our network-based gene prioritization approach provides a general framework for identifying and ranking candidate genes associated with complex diseases. To our knowledge this is the first time that protein interactions, ortholog mapping, gene expression, and literature mining were integrated for ranking candidate genes in any livestock species. The approaches and techniques that were systematically implemented in the present study are general and not confined by specific trait or species and can be applied to various complex diseases in different organisms.

GRANTS

The work was supported by Cutting Edge Genomics for Sustainable Animal Breeding (SABRE) research project that has been cofinanced by the European commission within the 6th Framework Programme, contract no. FOOD-CT-2006-016250. Additional financial support was kindly granted by Quantomics, a collaborative project under the 7th Framework Programme (FP7).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS


REFERENCES

3. Buitenhuis B, Rontved CM, Edwards SM, Ingvartsen KL, Sorensen P. Gene expression, and literature mining were integrated for this is the first time that protein interactions, ortholog mapping, aggregation and visualisation methods.