A Transcriptome-Interactome Network Of Protein Complexes Implicated In Bovine *E.coli* Mastitis

P. Sørensen[*], A. Skarman[*], L. Jiang[*], K. L. Ingvartsen[†] and C.M. Røntved[†]

Introduction

Most complex diseases including susceptibility to mastitis have a complex inheritance and may result from variants in many genes, each contributing only a small effect to the trait. We have successfully used linkage and linkage disequilibrium analyses to identify quantitative trait loci (QTL) affecting the resistance to clinical and subclinical mastitis (Sahana et al. (2008), Lund et al. (2008)). Further investigations showed that these QTL affect levels of genetic resistance specifically against the most common mastitis pathogens including *S. aureus* and *E. coli* (Sørensen et al. (2008)).

To gain further insight into the genetic background for mastitis-related traits we are now using two complementary high-throughput experimental approaches. The first approach is genome-wide association analysis (Balding (2006)) which is a powerful approach to identify regulatory genomic regions associated with the traits. These genomic regions may contain tens or even hundreds of genes. None the less identifying causative genes remains a challenge. The second approach is genome-wide expression analyses (Allison et al. (2006)) which on the other hand identifies changes in expression levels of genes and genetic pathways that associate with the trait. This approach does not provide information about the genetic variation underlying the trait. The challenge is that while these experimental approaches do provide a list of genes which are useful for classification and characterization of the disease, they do not provide sufficient information to identify specific causal genes.

Disease prediction methods exist for generating distinct prioritizations of candidate genes for human diseases (Aerts at al. (2006), Rossi et al. (2006), Lage et al. (2007), Chen et al. (2007), Wanderer et al. (2007)). These methods are typically based on the similarity of disease gene

[*] Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, Aarhus University, DK-8830 Tjele, Denmark

[†] Department of Animal Health and Biosciences, Faculty of Agricultural Sciences, Aarhus University, DK-8830 Tjele, Denmark

characteristics (e.g. sequence features, functional annotations, expression patterns and literature descriptions). Disease prediction methods based on protein-protein interaction networks (interactome) have recently been proposed based on the assumption that functionally related genes or gene products may be involved in the same or similar phenotypes. The central idea in these network based methods is that mutations in different members of a gene/protein complex may lead to similar diseases (Lage et al. (2007)). Thus once a complex having members involved in one disease has been identified, the rest of the members become candidates for having a biological relationship with the same disease. These methods differ in the sources of data and the methods used for combining them and to our knowledge they have not been tested on any livestock species including bovine.

We reasoned that functional network based inference is also suitable for genome-wide expression (transcriptome) data. Therefore the aim of this study is to predict and rank genes involved in the acute phase response to *E.coli* in the mammary gland of Danish Holstein cattle through the integration of protein-protein interactions and genome-wide expression profiles.

Material and methods

Genome wide expression data

Animals and Treatment. Sixteen, healthy primiparous Danish Holstein-Friesian cows were challenged intra mammarily with E.coli (k2bh2) 4 to 6 weeks after parturition. Quarters for *E.coli* inoculation and biopsy were selected based on CMT scores (≤ 2) and SCC in fore milk using the portable DeLaval Cell Counter (DCC; DeLaval, Tumba, Sweden) (range 1-6000 x 10^3 cells/ml). The front quarter with the lowest SCC (< 27,000 cells /ml) was chosen for E.coli inoculations. Control quarters where determined on the bacteriological examinations conducted prior to *E.coli* inoculation and on the quarter fore milk SCC at 24 h (< 181.000 cells /ml). The cows were housed in a traditional straw-bedded tiestall barn, where they were fed individually, and with free access to water. A total mixed ration (TMR) diet including vitamins and minerals was fed *ad libitum* twice a day in equal portions at 8.00 a.m. and 15.30 p.m. The cows were milked at 6.00 a.m. and again at 17.00 p.m. All procedures involving animals were approved by the Danish Animal Experiments Inspectorate and complied with the Danish Ministry of Justice Laws concerning animal experimentation and care of experimental animals

Udder biopsies. 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. The udder biopsies were frozen immediately in liquid nitrogen and transported to the laboratory where the tissues were stored at −80°C until RNA extraction.

RNA extraction and gene expression profiling. Isolation and labeling of RNA and microarray processing were performed as described elsewhere (Kristensen et al. (2005)). In total 21 samples were used (12 infected quarter and 9 control quarter). The expression profiles were measured using the Affymetrix Bovine Genome array (Affymetrix Clara CA, USA). The array contains 24,128 probe sets.

Annotation and statistical analysis. The annotation and analysis was performed using R version 2.10.0 (http://www.r-project.org/). Normalization of the arrays was performed using the robust multi array analysis (RMA) algorithm (Irizarry et al. (2003)). Differential expression between treatment groups (24h infected versus 24h control) was assessed using linear modelling and empirical Bayes methods as implemented in LIMMA (Smyth (2004)). Annotations for probes and bovine entrez genes were collected from the org.Bt.eg.db, org.Hs.eg.db and hom.Hs.inp.db annotation packages (www.bioconductor.org).

Gene ranking method

All known bovine entrez genes were considered as potential candidate genes for being implicated in bovine mastitis. We used an integrated approach to identify candidate gene complexes from protein-protein interactions and to rank the candidate gene based on the expression profile of its complex.

Candidate gene complexes were identified from protein-protein interactions gathered from the STRING (Jensen et al. (2009)) database. STRING is a comprehensive dataset containing functional links between proteins on the basis of both experimental evidence for protein-protein interactions as well as interactions predicted by comparative genomics and text mining. The protein(s) encoded by the candidate gene was used to retrieve first-order protein interaction partners. Since protein-protein interactions often are evolutionarily conserved, we used interaction data from bovine and from human based on homology. The transcriptome profile of each candidate complex was based on experimental data from a comprehensive genome-wide expression study of the host-pathogen response in cows described above.

For each candidate gene we then computed a score based on the transcriptome profile of its candidate complex. Our scoring model is based on a random-set scoring method currently used in gene set analyses (Newton et al. (2007)). This method use predefined sets of genes (e.g. candidate complex) and individual gene measures (e.g. log-fold change or t statistics). For each candidate gene complex a z-score is calculated and the relative importance of each of the candidate genes is based on the rank of the z-score.

We tested for enrichment of Gene Ontology terms among the 100 highest ranking genes using the functional annotation clustering approach implemented in the DAVID bioinformatics database (Dennis et al. (2003)).

Results and discussion

We have used protein-protein interactions and genome-wide expression profiles to predict and rank genes involved in bovine *E.coli* mastitis. The high ranking candidates includes well known immune genes such as Tumor Necrosis Factor-alpha (TNF-α), Lacto(trans)ferrin (LTF) and Toll-like receptor 4 (TLR-4) (Table 1). Test for enrichment of Gene Ontology terms among the 100 highest ranking genes showed a strong enrichment ($p<10^{-10}$) of genes involved in inflammatory, immune and defense responses.

Table 1. The 10 highest ranking genes by combined z-score

Symbol	logFC	Z_{Bt}	Z_{Hs}	Z_C	R_{Bt}	R_{Hs}	Gene Name
TNF-alpha	0.21	6.97	9.48	8.22	19	1	tumor necrosis factor (TNF superfamily, member 2)
LTF	0.96	7.15	6.70	6.93	16	7	Lacto(trans)ferrin
TLR4	0.76	8.94	4.46	6.70	4	82	toll-like receptor 4
NCF1	3.20	9.58	3.77	6.67	2	162	neutrophil cytosolic factor 1
S100A9	6.22	9.24	4.08	6.66	3	120	S100 calcium binding protein A9
S100A8	6.37	8.81	3.74	6.28	5	173	S100 calcium binding protein A8
IL1B	1.32	10.28	2.27	6.27	1	626	interleukin 1, beta
IL4	0.05	6.88	5.31	6.10	23	31	interleukin 4
MS4A8B	1.84	3.58	8.42	6.00	289	2	membrane-spanning 4-domains, subfamily A, member 8B
HMBOX1	0.37	6.48	5.48	5.98	30	22	homeobox containing 1

logFC: log fold change, Z_{Bt} (R_{Bt}) is the bovine transcriptome z-score (rank), Z_{Hs} (R_{Hs}) is the human transcriptome z-score (rank).

The Z-score computed above is based on expression changes of candidate complexes in response to *E.coli* infection. It does not provide any evidence of causality, but it is useful for prioritization of candidate genes located within genomic regions found to be associated with clinical mastitis. To infer causality it is necessary to perform an integrated analysis of genotypic data, gene expression data and clinical data (Brem et al. (2002), Schadt et al. (2005)). This could enable one to study DNA variations in the genome and the perturbations these variations give rise to at the candidate complex expression level, which in turn lead to resistance to clinical mastitis.

Our transcriptome-interactome approach relies on publicly available data including protein-protein interactions (e.g. STRING), but it is important to notice that our approach is not limited to genes with protein interactions. As long as the gene itself has one or more gene-associated phenotypes we can quantify the association between the gene and the disease phenotype. This is a clear advantage as compared to existing network based disease prediction approaches.

Our transcriptome-interactome approach can be improved by integrating additional sources of gene associated phenotypes. First, genome-wide expression profiles from the udder of *Staphylococcus aureu* (Strandberg Lutzow et al. (2008)), *Streptococcus uberis* (Moyes et al. (2009)), and *E.coli* (Rinaldi et al. (2010)) infections is publicly available. If we include data from various studies of host-pathogen responses we will gain further insights into pathogen specific host transcriptional responses of candidate complexes. Second, the phenome-interactome approach used for disease predictions in humans can be adapted for use in livestock species (Lage et al. (2007)). In this approach a phenome profile of each candidate complex is determined. This involves linking genes to phenotypes and determining similarity measures of phenotypes. Links between genes and phenotypes exists in publicly available databases such as Online Mendelian Inheritance in Animals (Nicholas (2003)) and from similar resources for humans (Hamosh et al. (2002)). Similarity measures of phenotypes are based on text-mining of a textual description of diseases and complex traits found in these databases (Lage et al. (2007)). We are currently testing a phenome-interactome approach which we have adapted for livestock species. Our goal is to integrate these data into our approach and thereby improve the prioritizations of candidate genes for complex diseases in livestock species.

Conclusion

In this study we have outlined a disease gene prediction approach for identifying and ranking genes associated with other complex diseases in livestock populations. Especially as more information about gene associated phenotypes and protein-protein interaction data becomes available, we expect that our network based disease prediction approach will become ever more accurate in its ranking of candidate genes.

References

Sahana, G., et al. (2008). *Anim. Genet.* 39(4):354-62.
Lund, MS., et al. (2008). *J. Dairy. Sci.* 91(10):4028-36.
Sørensen, LP., et al. (2008). *J. Dairy. Sci.* 91(6):2493-500.
Balding, DJ. (2006). *Nat. Rev. Genet.* 7(10):781-91.
Allison, DB., et al. (2006). *Nat. Rev. Genet.* 7(1):55-65.
Aerts, S., et al. (2006). *Nat. Biotechnol.* 24(5):537-44.
Lage, K., et al. (2007). *Nat. Biotechnol.* 25(3):309-16.
Chen, J., et al. (2007). *BMC Bioinformatics.* 8:392.
Rossi, S., et al. (2006). *Nucleic Acids Res.* 34:W285-W292.
Wanderer, AA., et al. (2008). Am. J. *Hum. Genet.* 82:949-958.
Kristensen, TN., et al. (2005). *Genetics.* 171:157-167.
Irizarry, RA., et al. (2003). *Biostatistics.* 4: 249-264.
Smyth, GK. (2004). *Stat. Appl. Genet. Mol. Biol.* 3(1):Article 3.
Jensen, LJ., et al. (2009). *Nucleic Acids Res*. 37:D412-6.
Newton, MA., et al. (2007). *Annals of Applied Statistics*, 1(1):85-106.
Dennis, G., et al. (2003). *Genome. Biol.* 4(5):P3.
Schadt EE. et al. (2005). *Nat. Genet.* 37:710-717.
Brem, RB., et al. (2002). *Science.* 296:752-755.
Strandberg Lutzow, YC., et al. (2008). *BMC Vet. Res.* 4:18.
Moyes, KM., et al. (2009). *BMC Genomics*, 10:542.
Rinaldi, M., et al. (2010). *Funct. Integr. Genomics*, 10:21-37.
Nicholas, FW. (2003). *Nucleic Acids Res*. 31:275-277.
Hamosh, A., et al. (2002). *Nucleic Acids Res*. 30:52-55.