University of Veterinary Medicine Hannover

Institute for Animal Breeding and Genetics

Bioinformatical Meta-Analysis of High-Throughput Expression Data from Neuroinfection Research

THESIS

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Natural Sciences

Doctor rerum naturalium

(Dr. rer. nat.)

awarded by the University of Veterinary Medicine Hannover

by

Robin Kosch

from Höxter

Hannover, Germany 2019

Supervisor: Prof. Dr. Klaus Jung

Supervision Group: Prof. Dr. Klaus Jung
Prof. Dr. Stefanie Becker
PD Dr. Martin Eisenacher

1st Evaluation: Prof. Dr. rer. nat. Klaus Jung
Institute for Animal Breeding and Genetics
University of Veterinary Medicine Hannover, Foundation

Prof. Dr. rer. nat. Stefanie Becker
Institute for Parasitology (Dept. of Infectious Diseases)
University of Veterinary Medicine Hannover, Foundation

PD Dr. rer. medic. Martin Eisenacher
Medical Proteome-Center
Ruhr-University Bochum

2nd Evaluation: Prof. Dr. rer. nat. Anne-Laure Boulesteix
Institut für Medizinische Informationsverarbeitung,
Biometrie und Epidemiologie (IBE)
Ludwig-Maximilians-Universität München

Date of final exam: 03.04.2019

Parts of the thesis have been published previously in:

Kosch, R., Delarocque, J., Claus, P., Becker, S. C., & Jung, K. (2018). Gene expression profiles in neurological tissues during West Nile virus infection: a critical meta-analysis. *BMC Genomics*, *19*(1), 530.

Kosch, R. & Jung, K. (2019). Conducting Gene Set Tests in Meta-Analyses of Transcriptome Expression Data. *Research Synthesis Methods, 10(1), 99-112.*

Sponsorship:

This study was supported by the Niedersachsen-Research Network on Neuroinfectiology (N-RENNT) of the Ministry of Science and Culture of Lower Saxony.

Table of Contents

Abbreviations

cDNA	-	complementary DNA
DE	-	differentially expressed
DEG	-	differentially expressed gene
EMBL-EBI	-	European Molecular Biology Laboratory - European Bioinformatics Institute
FDR	-	false discovery rate
GEO	-	Gene Expression Omnibus
GO	-	gene ontology
GSEA	-	Gene set enrichment analysis
GT	-	Globaltest
MIAME	-	Minimum information about a microarray experiment
NCBI	-	National Center for Biotechnology Information
PRISMA	-	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RHD	-	RepeatedHighDim
RMA	-	Robust Multi-array Average
RNA-Seq	-	Ribonucleic acid-Sequencing
ROAST	-	Rotation Gene Set Tests
ROC	-	receiver operating characteristic
ROMER	-	Rotation testing using Mean Ranks
RT-qPCR	-	real-time quantitative polymerase chain reaction
WNV	-	West Nile virus
ZIKV	-	Zika virus

Summary

Bioinformatical Meta-Analysis of High-Throughput Expression Data from Neuroinfection Research

Robin Kosch

The measurement of gene expression levels by microarrays or next-generation sequencing techniques (RNA-Seq) allows researchers to examine a range of biological questions. Hence, the amount of such transcriptome data has increased dramatically within the last 20 years, by far not solely in the field of neuroinfection research. Their number, as well as their availability leads to a more frequent usage of meta-analyses. These enable to aggregate the findings of individual studies, which can result in an enhanced overall statistical power and a higher degree of scientific evidence.

Classical meta-analyses are usually conducted by pooling the results of individual studies. As an alternative approach, it is possible to merge the studies directly on the data level. Hence, two data integration pipelines for meta-analyses were tested in this thesis: the 'early merging' approach with combination of raw transcriptome data from multiple studies and the 'late merging' by combining their individual results.

While current bioinformatical methods for meta-analyses mainly focus on standard differential expression analyses, other concepts typically performed in individual data sets were rarely considered in meta-analysis. Among these concepts are for example gene set enrichment and global test approaches to analyze not individual genes, but sets of genes. Analyzing gene sets is an important step for better understanding the biological impact of a treatment or disease. Therefore, the usage of several statistical tests for gene sets in meta-analysis were investigated.

In the first publication of this thesis, an applied example of a meta-analysis with real biological expression data was conducted. Thereby, the different integration methods were investigated by applying the two analysis pipelines ('early' and 'late merging'). Initially, public repositories for studies from neuroinfection research were screened and suitable gene expression datasets were found for West Nile virus infected mice. Five of these datasets could be related to neurological tissues (including 44 samples in total), whereas two datasets were related to immunological tissues (including 18 samples in total). The meta-analysis of each group allowed identifying differentially

expressed genes that were not identified by the individual studies alone. Further, lists of enriched gene sets, defined by gene ontology terms were revealed. From the overall top20 DE genes of the neurological tissues, eight genes commonly appeared in both analysis pipelines. These genes were discussed in a biological context and their antiviral activity and their participation in interferon cell signaling pathways could be confirmed, all correlated with the WNV-infection. Besides these biological investigations, publication I provides a practical example of an early stage data integration method for a meta-analysis, including the step of gene set analysis.

In the second publication of this thesis, these methodical findings were extended, by meta-analyses of simulated transcriptome data. For that purpose, data matrices were created by drawing expression values from multivariate normal distributions. To simulate heterogeneous conditions, batch effects were added to the artificial datasets in multiple ways. A more realistic, but less controllable approach to simulate data was conducted by utilizing a real existing dataset and dividing it into several subgroups that were considered to represent independent studies. For both approaches, the gene sets for the pathway analysis were simulated as well. Thereby, a predefined outcome could be generated, on which the methods were tested. The flexibility of the simulated data allowed varying the study size and the level of heterogeneity. Thus, an overall higher sensitivity of the 'early merging' strategy to detect enriched gene sets could be exposed compared to the 'late merging'. Only for simulation scenarios with fewer studies, but larger sample sizes and large batch effects, the 'late merging' strategy has been shown to be superior. Conclusively, the choice of the strategy is still highly based on the study and sample sizes. The heterogeneity between the datasets has been also shown as an essential factor. Competitive approaches were exposed as a practical method, a lot more than the self-contained methods. ROMER showed the highest sensitivity, but might lack in accuracy. Therefore, GSEA by Subramanian et al. appeared to be a good choice.

Further, the credibility and performance, but also limitations of simulation scenarios for meta-analyses of transcriptome data were unveiled.

Zusammenfassung

Bioinformatische Meta-Analyse von Hochdurchsatz-Expressionsdaten von Neuroinfektionsstudien

Robin Kosch

Die Messung von Genexpressions-Leveln mit Hilfe von Microarrays oder Next-Generation Sequencing Techniken (RNA-Seq) ermöglicht es Forschern eine Vielzahl an biologischen Fragestellungen zu untersuchen. Dadurch hat die Menge an Transkriptomdaten innerhalb der letzten 20 Jahre, nicht nur im Bereich der Neuroinfektionsforschung, enorm zugenommen. Sowohl die Anzahl an Datensätzen, als auch ihre Verfügbarkeit führt zur vermehrten Verwendung von Metaanalysen. Diese ermöglichen die Synthese von Ergebnissen aus individuellen Studien, was in einer erhöhten statistischen Power, sowie in einem höheren Level an wissenschaftlicher Evidenz resultieren kann.

Klassische Metaanalysen werden üblicherweise durchgeführt, indem die Ergebnisse von Einzelstudien vereinigt werden. Als Alternative ist es möglich, die Studien bereits auf Datenebene zusammenzufügen. Daher wurden zwei Integrations-Pipelines für Metaanalysen untersucht: die „early merging"-Strategie zur Aggregation der Transkriptom-Rohdaten aus mehreren Studien, sowie die „late merging"-Strategie zur Synthese der Ergebnisse aus Einzelstudien.

Während aktuelle bioinformatische Methoden für Metaanalysen standardmäßige differentielle Expressionsanalysen thematisieren, werden andere Konzepte, die typischerweise auf individuellen Datensätzen angewendet werden in Metaanalysen nur wenig berücksichtigt. Als Beispiele lassen sich hier die „Gene Set Enrichment Analysen" oder Globaltests nennen, welche nicht auf die Analyse von Einzelgenen, sondern Gen-Gruppen abzielen. Die Analyse solcher Gen-Gruppen ist ein elementarer Bestandteil um den biologischen Einfluss einer Behandlung oder Krankheit zu verstehen. Daher wurden mehrere statistische Herangehensweisen für Gen-Gruppen in Metaanalysen getestet.

Die erste Publikation dieser Arbeit zeigt ein angewandtes Beispiel einer Metaanalyse mit realen biologischen Expressionsdaten. Dabei wurden die unterschiedlichen Integrationsmethoden

innerhalb der Analyse-Pipelines (‘early und ‘late merging’) getestet. Zunächst wurden öffentliche Datenbanken nach Studien aus der Neuroinfektionsforschung durchsucht. Expressionsdatensätze zu West Nil Virus-infizierten Mäusen stellten sich als geeignet heraus. Fünf von diesen Datensätzen konnten neurologischen Geweben zugeordnet werden (insgesamt 44 Samples), wohingegen zwei Datensätze immunologischen Geweben (insgesamt 18 Samples) zugeordnet werden konnten. Durch eine Metaanalyse der jeweiligen Gruppen wurden signifikant differentiell exprimierte Gene identifiziert, welche in den Ergebnissen der Einzelstudien nicht detektiert worden sind. Außerdem konnten Listen von differentiell überrepräsentierten Gen-Gruppen nach Klassifizierung der Gene Ontology-Terminierung erstellt werden. Aus den Top20-Genen der neurologischen Gewebe beider Pipelines wurden acht Gene gemeinsam detektiert. Diese Gene wurden in ihrem biologischen Kontext diskutiert, wodurch ihre antiviralen Eigenschaften und ihre Teilnahme an Interferon-Zell-Signalwegen bestätigt werden konnten, korreliert mit der WNV-Infektion. Neben diesen biologischen Erkenntnissen, konnte ein praktisches Beispiel der initialen Datenintegration für eine Metaanalyse inklusive einer Gene Set Analyse erfolgreich dargestellt werden.

In der zweiten Publikation dieser Arbeit wurde die Herangehensweise um eine Metaanalyse mit simulierten Transkriptomdaten erweitert. Dazu wurden Datensätze erstellt mit Expressionswerten, gezogen aus multivariaten Normalverteilungen. Um heterogene Bedingungen zu generieren, wurden zusätzlich Batch-Effekte in verschiedenen Ausführungen auf die künstlichen Datensätze hinzugerechnet. Ein realistischerer, jedoch weniger kontrollierbarer Ansatz einer Simulationsstudie wurde durchgeführt, indem ein realer Datensatz in mehrere Subdatensätze aufgeteilt wurde, welche unabhängige Einzelstudien repräsentieren sollen. Für beide Ansätze wurden die Gen-Gruppen für die Pathway-Analyse ebenfalls simuliert. Somit konnte ein bereits bekanntes Ergebnis generiert werden, um so die Methoden zu überprüfen. Die Flexibilität der simulierten Daten ermöglichte es, die Größe und Heterogenität der Studien zu variieren. So konnte für die „early merging“-Strategie eine insgesamt höhere Sensitivität bei der Detektion von differentiell exprimierten Gen-Gruppen festgestellt werden im Vergleich zur „late merging“-Strategie. Nur für Simulationsszenarien mit wenig Studien, vielen Samples und größeren Batch Effekten stellte sich das „late merging“ als vorteilhafter heraus. Abschließend ist die Wahl der Integrationsmethode stark abhängig von Studien- und Samplegröße. Auch die Heterogenität zwischen den Datensätzen ist ein essentieller Faktor. Kompetitive statistische Herangehensweisen

erwiesen sich als deutlich praktikabler als in sich abgeschlossene (engl. self-contained) Tests. ROMER zeigte die höchste Sensitivität, ist jedoch möglicherweise weniger präzise. Daher erscheint GSEA von Subramanian et al. eine gute Wahl der Methode.

Weiterhin konnten die Durchführbarkeit und Performance, aber auch die Grenzen von Simulationsstudien für Metaanalysen von Transkriptomdaten aufgezeigt werden.

1. Introduction

1.1. Classical meta-analyses

Meta-analyses are widely used in clinical or epidemiological trials to integrate the outcome of multiple studies regarding one specific biological or medical question. Therefore, the individual results are summarized or statistically reanalyzed as one single study. Such meta-analyses, sometimes also referred to as ‘research syntheses’, can be part of systematic reviews, but extend those by a quantitative analysis. The term ‘meta-analysis’ was coined by Glass (1976) in the field of educational research.

By increasing the number of observations, meta-analyses result in higher statistical power, can reduce bias and allow more precise interpretations of the study question in contrast to the analysis of individual studies. The number of false positive results can be reduced as well. Further, contradictory outcomes can be detected more easily and misleading conclusions appear therefore less frequently. Another major task of meta-analyses is the examination of heterogeneity between study results (Haidich 2010). Thus, meta-analyses are a substantial technique for evidence-based medicine.

In the regard of the ‘reproducibility crisis’ that has increasingly been discussed in the recent years (Ioannidis 2005), meta-analysis can also provide a tool to bring more robustness and reproducibility into research, by enhancing sample sizes (Baker 2016). Moreover, non-biological effects, e.g. caused by different lab conditions can be reduced.

1.2. Meta-analyses of transcriptome data

Meta-analyses can be performed with various types of input data. This thesis covers the field of transcriptomics. A great amount of available data is still generated by cDNA microarrays, whereas these will be replaced by newer technologies, e.g. RNA sequencing. Nevertheless, the here presented findings may be cautiously transferable to diverse types of high-throughput expression data, such as proteomics or metabolomics.

Transcriptome expression analysis allows researchers to study biological processes, developments or diseases, all on a genetic level. Further microarrays can be used for diagnostic purposes or for the prediction of therapy response. This will be achieved by quantifying the transcripts of subjects

under different conditions. A brief historical overview of DNA microarrays in biomedicine can be obtained by Ewis et al. (2005).

A crucial difference between clinical and transcriptome data lie in their dimensionality. Whereas clinical studies only focus on few variables studied on relatively large sample sizes (number of samples > number of studied patient characteristics), transcriptome analyses produce high-dimensional data with ten thousands of features or more. Nevertheless, the number of samples within gene expression analyses is usually low (number of genes > number of samples), which makes them susceptible for incorrect interpretations, due to lacks of scientific evidence.

A basic step of a meta-analysis is the selection of suitable studies, i.e. with the same design and a comparable study question. For transcriptome experiments, it is common that researchers publish not only their findings as part of publications, but also deposit their raw expression data in public databases. There are two well-known public repositories for gene expression data: Gene Expression Omnibus (GEO) from NCBI (Edgar 2002) and ArrayExpress from EMBL-EBI (Brazma 2003). Additionally, the submitted data will be at least partly curated by a team of experts. Both archives also provide web-based tools for the further analysis. These beneficial sources of expression data can be the starting point of meta-analysis. For creating overview, the repositories are screened intensively, depending on the initial biological or medical questions. To get a more comprehensive view, the entire literature will be searched manually.

The availability of raw high-throughput gene expression data makes it possible to reproduce the author´s findings, but also to extend them by a joint analysis of similar studies. Further, these repositories support the annotation standards, called MIAME (Brazma et al. 2001), which are necessary for an appropriate description of the transcriptome experiments.

The percentage of published studies related to the term ‘meta-analysis’ has been increased enormously within the last 30 years. The same situation can be observed for publications, focusing on ‘transcriptomic’ experiments, whereas these were introduced by the technological improvements since 1999. Nevertheless, the amount of studies, which cover both terms, meta-analysis and transcriptome data is still small (Brown and Peirson 2018).

Many examples of a successful meta-analysis of transcriptomic data are available for very diverse research fields. Te Pas et al. (2012) examined different chicken lines regarding their susceptibility to Salmonella infection, whereas Desterke et al. (2018) highlights the importance of trophoblastic

differentiation in hydatidiform mole. Another contrary meta-analysis application was proposed by Balan et al. (2018), who analyzed the transcriptome response to biotic stresses in apple (Malus x domestica). Specific meta-analyses of neuroinfection studies can be obtained from Afroz et al. (2016). In this publication the gene expression data of several studies with Dengue virus infections were aggregated, which revealed novel gene signatures, caused by the pathogen. In a further RT-qPCR, the findings of the meta-analysis have been validated. Another representative of the flaviviruses was the research object of Singh et al. (2018). They examined Zika virus-induced (ZIKV) expression profile changes by utilizing multiple datasets, but also compared them to other related pathogens (Japanese encephalitis, West Nile, and Dengue). In their meta-analysis, a characteristic ZIKV infection signature was identified.

While meta-analyses are becoming more and more popular, the majority of already existing valuable transcriptome expression data is still untapped.

1.3. Goals of the study

By conducting multiple meta-analysis of real biological transcriptome expression data (publication 1), as well as on simulation data (publication 2), this thesis aimed to test an alternative way of data merging within the context of gene set enrichment analysis. First, the feasibility of the new analysis pipeline was examined; second, it was analyzed how the structure and size of the aggregated studies influence the final outcome. Further, a comparison of two different strategies of pathway analysis was conducted: competitive and self-contained methods. Therefore, multiple statistical approaches were used within each strategy. Another objective of doing such meta-analysis was the impact of heterogeneity between the studies, but also between the samples within studies. This issue was tackled, by introducing diverse artificial batch effects on simulated data and compare their results. Such meta-analyses of transcriptome data from WNV-infected patients have not been conducted before.

The two data integration strategies for meta-analyses were evaluated regarding their capability of detecting significantly enriched gene sets by performing simulation studies. Therefore, three different gene set enrichment tests were utilized, as well the three different approaches for the globaltest strategy, which have not been explored in this context. The flexible study design of artificial transcriptome data allowed examining the behavior of the analysis pipelines. Further, this study employs an approach of simulating the expression values with correlations between the

genes, whereas other studies simulated genes as uncorrelated. Different study sizes, but also sample sizes were tested to provide orientation data selection of future meta-analyses.

2. Materials and Methods

In this section, the data and methods that were used in publication 1 and 2 are briefly described and the reader is referred to more detailed descriptions in these publications.

2.1. Data selection and construction

In order to investigate the outcome of the two merging strategies, both were applied on multiple data origins. First, datasets from real biological high-throughput experiments were selected by screening the repositories for neuroinfection studies. The resulting tables allowed a broad view on the current (i.e. time of database screening) available transcriptome data on several pathogens. Due to the absence of many studies on some of the pathogen groups, the further selection was limited a lot. The most promising results were detected for the West Nile virus (Tab. 1; Petersen et al. (2013)). For the selection process, the PRISMA statement offered guidance, which among others provides exclusion criteria to promote the selection success (Moher et al. 2009).

Tab. 1: Overview of the datasets used in publication I and II.

	Publication I	**Publication II**
Real data	WNV-infection (neurological tissues) WNV-infection (immunological tissues)	Rhinovirus-infection (with simulated pathways) Prostate cancer data
Simulation data	none	Several simulation scenarios

The simulation of expression data was conducted in a common way by drawing values from a multivariate normal distribution (Ghaffari et al. 2013; Hua et al. 2009). Samples from the uninfected group were simulated by setting the mean vector for the multivariate normal distribution to null, whereas for infected samples the values were drawn from a univariate normal distribution in order to introduce the treatment effect. For the covariance matrix of both groups, an autoregressive model was used to simulate the correlation between the genes. Additionally, artificial batch effects were introduced by adapting the model from Johnson et al. (2007). Further,

gene sets were simulated with several degrees of enrichment. Table 1 provides an overview of all datasets within the two presented publications.

An alternative approach for simulating gene expression data has been shown by Ritchie et al. (2006).

2.2. Variants of meta-analysis

The classical way of conducting a meta-analysis is by aggregating the results of the individual studies. For data from gene expression analyses, this is typically conducted by p-value combination methods or combination of fold change estimates. An alternative way of conducting a meta-analysis can be obtained by aggregating the individual studies directly on the data level. In the following, the strategies were named for the sake of convenience ‘late merging’ and ‘early merging’ (Fig.1).

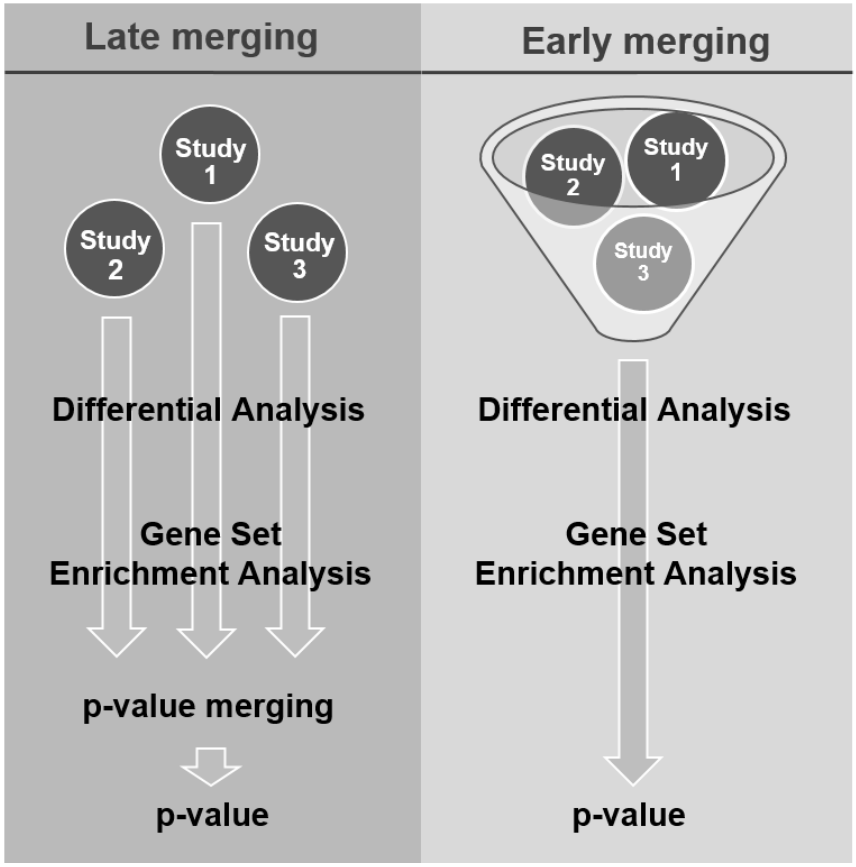

Fig. 1: Merging strategies for data integration

2.3. Steps of analysis pipeline

The analysis of high-throughput gene expression data can be conducted with a plethora of bioinformatical tools. For data simulation and manipulation, the statistical software R provides a suitable environment, which can be used under the GNU General Public License (Team 2018). Moreover, many beneficial software packages are available for R, especially from the open source software framework Bioconductor (Gentleman et al. 2004).

2.3.1. Preprocessing

Before the datasets can be merged for a meta-analysis, a couple of preprocessing steps are necessary. First, the genes from each array need to be annotated by the same gene identifiers. Next, the datasets are normalized and summarized, e. g. by Robust Multi-array Average (RMA) method, which is a widely used approach (Waldron and Riester 2016). Normalizing the data is an elementary to remove systematic biases within a study and therefore to make the samples comparable.

2.3.2. Batch effect removal

Selecting suitable studies for meta-analysis is an elementary step to produce proper results. Nevertheless, the comparison and joint analysis of data from multiple sources needs to be examined for their heterogeneity between individual studies. Since only then the unwanted study bias can be reduced before the analysis.

In order to perform a proper data synthesis, the employed studies need to be cleaned from batch effects. Therefore, biases caused by non-biological incidents are detected and widely minimized. A critical step during the removal is the preservation of biological effects. However, the complete removal of batch effects can never be achieved for practical meta-analysis of high-dimensional data (Waldron and Riester 2016).

For the analyses of within this thesis, the batch effect removal was only conducted in the ‘early merging’ pipeline. We selected the Bayesian method ‘ComBat’ (Johnson et al. 2007) from the R-package ‘sva’, which performed best in several tests (Chen et al. 2011).

2.3.3. Differential analysis

A typical goal of high-throughput gene expression studies is to detect genes that are differentially expressed between two conditions (Tusher et al. 2001; Smyth 2004). These can be different

developmental stages or experimental factors. This thesis focused on infected samples in comparison to non-infected control samples. That purpose can be achieved by testing multiple hypothesis, i.e. a gene-wise comparison of their gene expressions under both conditions.

For the here presented analysis pipelines, the popular 'limma'-method was used (Smyth 2004; Ritchie et al. 2015), which has been exposed as standard tool for microarray analysis. Limma allows the comprehensive investigation of the entire array by utilizing linear models on the genes. Further one can benefit from the 'high parallel nature of genomic data to borrow strength between the gene-wise models' (Ritchie et al. 2015). These tests result in large lists of p-values for each gene, representing their validity of being differentially expressed between the conditions.

Due to the huge amount of independent tests, high rates of false positives might be obtained. Therefore, the results were adjusted by controlling the false discovery rate (FDR) as proposed by Benjamini and Hochberg (1995).

2.3.4. Gene Set Enrichment Analysis

In order to enhance the benefit from transcriptome studies, genes can be classified into sets of functional genes. Hereby, genes that share the same biological pathway, chromosomal location or regulation are grouped and analyzed commonly. For our analysis, the gene ontology (GO) terms provided by the 'GO Consortium' (Ashburner et al. 2000; Harris et al. 2004) were utilized. For GO term classification, the genes were grouped within one of the following main categories: cellular components, molecular functions or biological processes. The gene sets were ordered hierarchically, from general pathways with multiple thousands of genes to very specific pathways with only few genes. Thus, the set-based analysis provides interactions between genes can be investigated, as well as an overall biological insight can be gained. In this work, the terms 'gene set' and 'pathway' are used as synonyms.

The first gene set enrichment analysis (GSEA) approach was introduced by Subramanian et al. (2005). The GSEA method works roughly as follows: first, an enrichment score is calculated by using the t-statistics of genes and an approach, similar to a Kolmogorov–Smirnov test for comparing genes within and outside a gene set. Additionally, a weighting is performed depending on the correlation of the genes to the phenotype. To determine the significance of an enrichment score for a gene set, the samples will be permutated and the score will be calculated again (Efron and Tibshirani 2007).

Besides the original GSEA approach, several other methods exist that focus on the detection of enriched gene sets in a competitive way (Luo et al. 2009). Furthermore, a plethora of gene set enrichment analysis comparisons has already been performed (Ackermann 2008; Ackermann and Strimmer 2009; Hung et al. 2012; Fridley et al. 2010; Nam and Kim 2008; Maciejewski 2014). Nevertheless, universally valid recommendations can still not be given.

In publication II, further GSEA methods were also investigated, i.e. the Wilcoxon rank-sum test (Wilcoxon 1945) and ROMER (Ritchie et al. 2015). Thus, only GSEA by Subramanian et al. was investigated in publication I, while in total three different methods for GSEA were evaluated in publication II (Tab. 2)

2.3.5. Competitive and self-contained tests

Concerning the null hypothesis, the gene set tests can be classified in competitive and self-contained methods (Goeman and Buhlmann 2007). While competitive methods test the H_0-assumption that genes within a set are not more DE than those genes not in the set, self-contained test assume that no genes in the set are DE. A direct comparison of the methods from the two classes is therefore not viable, due to their different methodical assumptions.

A popular example of self-contained tests is the Globaltest (GT) by Goeman et al. (2004). While competitive methods require lists of effect sizes as input, Globaltests consider the expression values of the genes only within the given gene set. For that approach, logistic regression models are utilized.

Tab. 2: Overview of the gene set detection methods used in publication I and II.

	Publication I	**Publication II**
GSEA	GSEA (Subramanian et al.)	GSEA (Subramanian et al.) Wilcoxon rank-sum test ROMER
Globaltest	n. a.	Globaltest (Goeman et al.) RepeatedHighDim ROAST

Further, RepeatedHighDim (Jung et al. 2011) and ROAST (Wu et al. 2010) are other self-contained tests, which have been compared to GT in publication II (Tab. 2). An advantage of the competitive test in contrast to self-contained approaches is that the influence of all genes on the array is considered.

3. Publication I

Gene expression profiles in neurological tissues during West Nile virus infection: a critical meta-analysis

Robin Kosch[1†], Julien Delarocque[1†], Peter Claus[2], Stefanie C. Becker[3, 4] and Klaus Jung[1, 4*]

[1] Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17p, 30559 Hanover, Germany

[2] Institute of Neuroanatomy and Cell Biology, Hannover Medical School, Carl-Neuberg-Str. 1, 30625, Hanover, Germany

[3] Institute for Parasitology, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17, 30559, Hanover, Germany

[4] Research Center for Emerging Infections and Zoonoses, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17, 30559 Hanover, Germany

[†] Robin Kosch and Julien Delarocque contributed equally to this work.

*Correspondence: klaus.jung@tiho-hannover.de

State of publication: published

Journal: BMC Genomics (date: 13 July 2018)

Available at: https://doi.org/10.1186/s12864-018-4914-4

Supplementary files of this publication can be obtained from the journal websites by following the upper link.

The extent of contribution from Robin Kosch to this article:

Performance of experiments: 80%

Analysis of experiments: 80%

Writing of the paper: 50%

Abstract:

Background: Infections with the West Nile virus (WNV) can attack neurological tissues in the host and alter gene expression levels therein. Several individual studies have analyzed these changes in the transcriptome based on measurements with DNA microarrays. Individual microarray studies produce a high-dimensional data structure with the number of studied genes exceeding the available sample size by far. Therefore, the level of scientific evidence of these studies is rather low and results can remain uncertain. Furthermore, the individual studies concentrate on different types of tissues or different time points after infection. A general statement regarding the transcriptional changes through WNV infection in neurological tissues is therefore hard to make. We screened public databases for transcriptome expression studies related to WNV infections and used different analysis pipelines to perform meta-analyses of these data with the goal of obtaining more stable results and increasing the level of evidence.

Results: We generated new lists of genes differentially expressed between WNV infected neurological tissues and control samples. A comparison with these genes to findings of a meta-analysis of immunological tissues is performed to figure out tissue-specific differences. While 5.879 genes were identified exclusively in the neurological tissues, 15 genes were found exclusively in the immunological tissues, and 44 genes were commonly detected in both tissues. Most findings of the original studies could be confirmed by the meta-analysis with a higher statistical power, but some genes and GO terms related to WNV were newly detected, too. In addition, we identified gene ontology terms related to certain infection processes, which are significantly enriched among the differentially expressed genes. In the neurological tissues, 17 gene ontology terms were found significantly different, and 2 terms in the immunological tissues.

Conclusions: A critical discussion of our findings shows benefits but also limitations of the meta-analytic approach. In summary, the produced gene lists, identified gene ontology terms and network reconstructions appear to be more reliable than the results from the individual studies. Our meta-analysis provides a basis for further research on the transcriptional mechanisms by WNV infections in neurological tissues.

4. Publication II

Conducting Gene Set Tests in Meta-Analyses of Transcriptome Expression Data

Robin Kosch[1] and Klaus Jung[1*]

[1] Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17p, 30559 Hanover, Germany

*Correspondence: klaus.jung@tiho-hannover.de

State of publication: published

Journal: Research Synthesis Methods (date: 27 December 2018)

Available at: https://doi.org/10.1002/jrsm.1337

Supplementary files of this publication can be obtained from the journal websites by following the upper link.

The extent of contribution from Robin Kosch to this article:

Performance of experiments: 90%

Analysis of experiments: 75%

Writing of the paper: 70%

Abstract:

Research synthesis, e.g. by meta-analysis, is more and more considered in the area of high-dimensional data from molecular research such as gene and protein expression data, especially because most studies and experiments are performed with very small sample sizes. In contrast to most clinical and epidemiological trials, raw data is often available for high-dimensional expression data. Therefore, direct data merging followed by a joint analysis of selected studies can be an alternative to meta-analysis by *p*-value or effect size merging, or more generally spoken, the merging of results.

While several methods for meta-analysis of differential expression studies have been proposed, meta-analysis of gene set tests have very rarely been considered, although gene set tests are standard in the analysis of individual gene expression studies. We compare in this work different strategies of research synthesis of gene set tests, in particularly the `early merging' of data cleaned from batch effects versus the `late merging' of individual results.

In simulation studies and in examples of manipulated real world data, we found that in most scenarios the early merging has a higher sensitivity of detecting a gene set enrichment than the late merging. However, in scenarios with few studies, large batch effect, moderate and large sample sizes late merging was more sensitive than early merging.

5. Discussion

The two publications, presented in this thesis mainly focus on the same overall goal: the examination of different integration pipelines for meta-analysis of transcriptome data. Real biological data, simulated data or simulated gene groups were utilized to examine the new analysis pipeline and a comparison to the classical meta-analysis approach was conducted.

The following part describes the findings of both publications and puts them in an overall context. In the first manuscript, the early merging pipeline was tested by conducting an applied meta-analysis of transcriptome neuroinfection data. Thus, not only an exemplary guidance is offered, but also biological improvements. Expression profiles of two types of mouse tissues infected by WNV were analyzed, which provided insights into the gene regulation. By now, several individual gene expression studies in neuroinfectional research have been performed (Qian et al. 2015; Bourgeois et al. 2011). Nevertheless, the current scientific knowledge lacked in suitable bioinformatical meta-analyses of neuroinfection data.

The second manuscript aimed to compare the two merging pipelines directly and to evaluate their performances on diverse data origins. Not only the GSEA method was tested, as proposed in publication I, but also two other competitive methods were investigated. Further, three Globaltest methods were examined, as different approaches on testing gene sets in a meta-analysis. Based on the result, advanced knowledge on the behavior of the early stage data integration method, induced by different input data is provided.

The findings of publication I mainly revealed an explicit profit of the early merging pipeline for practical meta-analyses. Nevertheless, a superiority over the ‘late merging’ pipeline could not be detected due to the study design. Solely publication II allows an evaluation of the pipelines regarding their detection capability and accuracy.

In this section, a comparison to studies similar to both manuscript is undertaken. Further, the validity of the analysis and the evaluation methods is discussed.

Some research has been already carried out, addressing the general concept of testing gene set analysis within the context of a meta-analysis. Shen and Tseng (2010) also tested several integration methods within a GSEA based pipeline. Their approaches are quite similar to the late stage integration and ‘intermediate merging’ strategy (Publ. I, Section ‘Meta-analysis’), but they did not consider the synthesis on data level. Further, the ‘late merging’ strategy differs from Shen

and Tsengs´ method MAPE_P. While publication I and II considered the intersection of genes over all studies, MAPE_P compared whole pathways, irrespectively of the total number of included genes. Thus, the data loss due to not matching genes in the same pathways across the studies is still present in the analyses within publication I. Nevertheless, the approaches of publication II seemed to be more meaningful for a method comparison.

Rosenberger et al. (2015) also provides methods for pooling the data from Gene set enrichment analysis, but within the context of genome-wide association studies. Hence, this cannot be directly compared to the findings within this work. Another major benefit of the simulation studies in this thesis is the usage of expression values with correlations between genes, drawn from multivariate instead of univariate normal distributions (Schäfer and Strimmer 2005).

Heterogeneity between studies can usually be assumed. However, the minimization of heterogeneity within meta-analyses is a well-studied, but still not solved issue. Chen et al. (2013) proposed a Bayesian method, which allows the simultaneous execution of the differential analysis and gene set enrichment analysis. Nevertheless, publication II addressed the problem by a practical approach.

A challenging task for evaluating the performance of the merging strategies and gene set tests was the selection of a proper measurement technique. The introduced approach to utilize the capability of identifying enriched genes only covers the true positive rates. Thus, the evaluation might not be profound, but was already applied successfully in other microarray studies (Wu et al. 2005). Moreover, to give a comprehensive insight into the method comparison, the rates of false positives were provided (Publ. II, Tab. 4). Further, receiver operating characteristic (ROC) curves were generated for some simulation scenarios, which illustrate the specificity against the sensitivity (Publ. II, Fig. A9, A10, A12-A15). This allowed the examination and evaluation of the pipelines´ accuracy. However, ROC curves are in principle only applicable for dichotomous cases. Consequently, we just could derive ROC curves for selected scenarios, comparing a non-enriched to a highly enriched pathway.

Besides the overall higher detection capability of the ‘early merging’ strategy, more accurate results may be obtained, due to the lower information loss of the initial data integration. While all expression values are processed in one analysis within the early stage integration, only summarized results are aggregated in the late stage integration approach. Nevertheless, a loss of biological

information is also present in the ‘early merging’ pipeline through the methods for batch effect removal.

The removal of batch effects between the studies is an inevitable step within the data preprocessing. Besides the ‘ComBat’ function, the performance of ‘removeBatchEffects’ from the ‘limma’-package was also tested, both on real datasets, which already have been analyzed (Tab. 1; Marot et al. (2009)). Nearly no differences could be detected between the results from the two batch effect removal methods. Therefore, the approach that the same model was used for adding batch effects to data and for removing the batch effects again was still reasonable. For other simulation scenarios with different types of batch effects, ‘ComBat’ performed also very well.

The overall good performance of the GSEA by Subramanian et al. was already stated in some publications (Maciejewski 2014), whereas other studies’ findings showing superiority of Goeman´s Globatest over GSEA (Tarca et al. 2013). Commonly, an extremely high sensitivity of the GT was observed by Tarca et al. and within publication II. However, the self-contained tests have been exposed as impractical detection methods in the analysis of this thesis.

In the following section, limitations of the here presented methods and emerging difficulties during the application of such meta-analysis are described

Selecting suitable studies for a proper meta-analysis is an elementary step. Therefore, depending on the research topic, the amount of studies, which can be merged for a data synthesis, is often rather small. The combination of information of five studies with neurological tissues and two studies with immunological tissues, which were separated in three individual datasets, was still possible as shown in publication I. For the meta-analysis of those datasets, the statistical power is certainly a lot higher compared to the analysis of individual studies.

The biological interpretation of the findings within this study could be extended with more effort, but the focus of this thesis relies in the method comparison. Nevertheless, these findings on the single gene-level are highly robust, but also conservative, due to the combination of the results from the ‘early’ and ‘late merging’ pipelines. Thus, the identification of those genes was unambiguously correct. The same applies for the gene set analysis results by combining the ‘early’, ‘intermediate’ and ‘late merging’ strategies (Publ. I, Fig A.5.3 & A5.5).

This work does not address to give precise recommendations for specific gene set test. Instead, it provides an insight into the behavior of meta-analyses with varying input data.

However, problems emerged frequently during the selection of the datasets from the systematic review. To pool the datasets in a reasonable manner, their experimental design has to be predominantly equal, for instance regarding to the analyzed organism, strain, tissue or cell type. Another challenge, which hampers the exploitation of available data, is the false description of provided experiments. During the selection process in publication I, several studies were found that show inconsistencies between their descriptions within the journal publication and the corresponding data uploads in the repositories. Further, some datasets contained abbreviations without any additional information. This makes the studies generally worthless for further analysis. Therefore, one could just apply to the providing researchers to upload their data in a properly and self-explaining and manner.

A major issue regarding the reproducibility of research findings was stated by Ioannidis (2005), who draw the attention to the high rates of false results and the low powered evidence of many study designs. Enhanced research standards or detailed method reports lead to results that are more accurate. However, those concepts are rather difficult to establish. Thus, increasing the sample sizes by aggregating multiple study data is a straightforward way to enhance the robustness of studies and generate better-evidenced results.

To extend the findings of the here presented meta-analysis approaches, the next paragraphs provide ideas for further analysis strategies or application fields.

For a standard meta-analysis only those genes are considered, which are covered by all individual datasets. To reduce this information loss while combining information across multiple datasets, a further approach is to take into account those genes that were not covered by all datasets, but were still removed due to the missing within one or few studies. Those genes could still be included for further analysis, but with slightly reduced power.

Besides the integration of expression data from multiple platforms, studies already exist, which combine data across the omics-type (Wu et al. 2012) or across different species (Fierro et al. 2008). These might also be an application field for future experiments with the innovative early merging strategy and gene set enrichment analysis. A broad overview of data integration methods of genomic data can be regarded from (Hamid et al. 2009).

The basic idea of merging data in a meta-analysis can be extended by different study selection approaches. Instead of screening for studies related to a specific pathogen, it may be meaningful to

collect studies with ‘only’ similar pathogens, but with the subjects showing the same symptoms. Therefore, a new data source can be exploited and more potential studies can be considered for a meta-analysis. This alternative comparison approach might also result in novel infection-related genetic patterns, which would have not been found by applying conventional methods.

6. References

Ackermann, M. 2008. A comparison of statistical methods for gene set enrichment analysis. In *Diploma thesis*. Technische Universität Dortmund: Department of Statistics.

Ackermann, M., and K. Strimmer. 2009. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10 (1):47.

Afroz, S., J. Giddaluru, M. M. Abbas, and N. Khan. 2016. Transcriptome meta-analysis reveals a dysregulation in extra cellular matrix and cell junction associated gene signatures during Dengue virus infection. *Sci Rep* 6:33752.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25 (1):25-29.

Baker, M. 2016. Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help. *Nature* 533 (7604):452-455.

Balan, B., F. P. Marra, T. Caruso, and F. Martinelli. 2018. Transcriptomic responses to biotic stresses in Malus x domestica: a meta-analysis study. *Sci Rep* 8 (1):1970.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*:289-300.

Bourgeois, M. A., N. D. Denslow, K. S. Seino, D. S. Barber, and M. T. Long. 2011. Gene expression analysis in the thalamus and cerebrum of horses experimentally infected with West Nile virus. *PLoS One* 6 (10):e24371.

Brazma, A. 2003. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* 31 (1):68-71.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29 (4):365-371.

Brown, L. A., and S. N. Peirson. 2018. Improving Reproducibility and Candidate Selection in Transcriptomics Using Meta-analysis. *J Exp Neurosci* 12:1179069518756296.

Chen, C., K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu. 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* 6 (2):e17238.

Chen, M., M. Zang, X. Wang, and G. Xiao. 2013. A powerful Bayesian meta-analysis method to integrate multiple gene set enrichment studies. *Bioinformatics* 29 (7):862-869.

Desterke, C., R. Slim, and J. J. Candelier. 2018. A bioinformatics transcriptome meta-analysis highlights the importance of trophoblast differentiation in the pathology of hydatidiform moles. *Placenta* 65:29-36.

Edgar, R. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 30 (1):207-210.

Efron, B., and R. Tibshirani. 2007. On Testing the Significance of Sets of Genes. *Annals of Applied Statistics* 1 (1):107-129.

Ewis, A. A., Z. Zhelev, R. Bakalova, S. Fukuoka, Y. Shinohara, M. Ishikawa, and Y. Baba. 2005. A history of microarrays in biomedicine. *Expert Rev Mol Diagn* 5 (3):315-328.

Fierro, A. C., F. Vandenbussche, K. Engelen, Y. Van de Peer, and K. Marchal. 2008. Meta Analysis of Gene Expression Data within and Across Species. *Curr Genomics* 9 (8):525-534.

Fridley, B. L., G. D. Jenkins, and J. M. Biernacka. 2010. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One* 5 (9).

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5 (10):R80.

Ghaffari, N., M. R. Yousefi, C. D. Johnson, I. Ivanov, and E. R. Dougherty. 2013. Modeling the next generation sequencing sample processing pipeline for the purposes of classification. *BMC Bioinformatics* 14:307.

Glass, G. V. 1976. Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher* 5 (10):3.

Goeman, J. J., and P. Buhlmann. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23 (8):980-987.

Goeman, J. J., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20.

Haidich, A. B. 2010. Meta-analysis in medical research. *Hippokratia* 14 (Suppl 1):29-37.

Hamid, J. S., P. Hu, N. M. Roslin, V. Ling, C. M. Greenwood, and J. Beyene. 2009. Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics* 2009.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, and C. Gene Ontology. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32 (Database issue):D258-261.

Hua, J., W. D. Tembe, and E. R. Dougherty. 2009. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition* 42 (3):409-424.

Hung, J. H., T. H. Yang, Z. Hu, Z. Weng, and C. DeLisi. 2012. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform* 13 (3):281-291.

Ioannidis, J. P. 2005. Why most published research findings are false. *PLoS Med* 2 (8):e124.

Johnson, W. E., C. Li, and A. Rabinovic. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8 (1):118-127.

Jung, K., B. Becker, E. Brunner, and T. Beissbarth. 2011. Comparison of global tests for functional gene sets in two-group designs and selection of potentially effect-causing genes. *Bioinformatics* 27 (10):1377-1383.

Luo, W., M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 10:161.

Maciejewski, H. 2014. Gene set analysis methods: statistical models and methodological differences. *Brief Bioinform* 15 (4):504-518.

Marot, G., J. L. Foulley, C. D. Mayer, and F. Jaffrezic. 2009. Moderated effect size and P-value combinations for microarray meta-analyses. *Bioinformatics* 25 (20):2692-2699.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 6 (7):e1000097.

Nam, D., and S. Kim. 2008. Gene-set approach for expression pattern analysis. *Brief Bioinform* 9.

Petersen, L. R., A. C. Brault, and R. S. Nasci. 2013. West Nile virus: review of the literature. *JAMA* 310 (3):308-315.

Qian, F., G. Goel, H. Meng, X. Wang, F. You, L. Devine, K. Raddassi, M. N. Garcia, K. O. Murray, C. R. Bolen, R. Gaujoux, S. S. Shen-Orr, D. Hafler, E. Fikrig, R. Xavier, S. H. Kleinstein, and R. R. Montgomery. 2015. Systems immunology reveals markers of susceptibility to West Nile virus infection. *Clin Vaccine Immunol* 22 (1):6-16.

Ritchie, M. E., D. Diyagama, J. Neilson, R. van Laar, A. Dobrovic, A. Holloway, and G. K. Smyth. 2006. Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics* 7:261.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43 (7):e47.

Rosenberger, A., S. Friedrichs, C. I. Amos, P. Brennan, G. Fehringer, J. Heinrich, R. J. Hung, T. Muley, M. Muller-Nurasyid, A. Risch, and H. Bickeboller. 2015. META-GSA: Combining Findings from Gene-Set Analyses across Several Genome-Wide Association Studies. *PLoS One* 10 (10):e0140179.

Schäfer, J., and K. Strimmer. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* 4.

Shen, K., and G. C. Tseng. 2010. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* 26 (10):1316-1323.

Singh, P. K., I. Khatri, A. Jha, C. D. Pretto, K. R. Spindler, V. Arumugaswami, S. Giri, A. Kumar, and M. K. Bhasin. 2018. Determination of system level alterations in host transcriptome due to Zika virus (ZIKV) Infection in retinal pigment epithelium. *Sci Rep* 8 (1):11209.

Smyth, G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102.

Tarca, A. L., G. Bhatti, and R. Romero. 2013. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One* 8 (11):e79217.

Te Pas, M. F., I. Hulsegge, D. Schokker, M. A. Smits, M. Fife, R. Zoorob, M. L. Endale, and J. M. Rebel. 2012. Meta-analysis of chicken--salmonella infection experiments. *BMC Genomics* 13:146.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Tusher, V. G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98.

Waldron, L., and M. Riester. 2016. Meta-Analysis in Gene Expression Studies. *Methods Mol Biol* 1418:161-176.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1 (6):80-83.

Wu, D., E. Lim, F. Vaillant, M. L. Asselin-Labat, J. E. Visvader, and G. K. Smyth. 2010. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* 26 (17):2176-2182.

Wu, S., Y. Xu, Z. Feng, X. Yang, X. Wang, and X. Gao. 2012. Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. *BMC Bioinformatics* 13:320.

Wu, W., N. Dave, G. C. Tseng, T. Richards, E. P. Xing, and N. Kaminski. 2005. Comparison of normalization methods for CodeLink Bioarray data. *BMC Bioinformatics* 6:309.

Danksagung

Ich bedanke mich sehr herzlich bei allen, die mir das Anfertigen dieser Arbeit in direkter aber auch in indirekter Weise ermöglicht haben.

Mein besonderer Dank gilt meinem Doktorvater Prof. Dr. Klaus Jung für die stets hervorragende Betreuung im gesamten Verlauf der Promotionszeit. Nur durch die vielen informativen Gespräche, entstandenen Ideen und kompetenten Ratschläge konnten die wissenschaftlichen Publikationen erstellt werden. Bedanken möchte ich mich auch dafür, dass Sie mir viele Freiheiten gewährt und mir viel Fairness entgegengebracht haben!

Für die weitere Betreuung und inhaltlichen Ratschläge bedanke ich mich auch bei meiner zweiten Betreuerin Prof. Dr. Stefanie Becker und externem Betreuer PD Dr. Martin Eisenacher.

Für die thematische Einarbeitung aber auch die wertvollen Tipps rund um die Dissertation möchte ich mich bei Dr. Jochen Kruppa bedanken. Als Bürokollege kann man sich kaum jemand besseren wünschen.

Für den Support in allen technischen Fragen bedanke ich mich bei Jörn Wrede.

Allen Mitarbeitern des Instituts, sowie meiner Arbeitsgruppe (Julien, Christine, Ihsan, Moritz) möchte ich für die tolle Zeit in der Tierzucht bedanken. Die erheiternden Kaffeepausen, Hunderunden und Mensagänge haben mir die letzten Jahre viel Freude bereitet. Wahrscheinlich gibt es kein Institut mit einer höheren Anzahl Kuchen pro Woche als die Tierzucht ☺

Für den Ablauf des PhD-Programms bedanke ich mich bei allen Organisatoren der HGNI.

Meinen Eltern, meinem Bruder und meiner Schwester danke ich für den bedingungslosen Rückhalt.