Viktor Jonsson Statistical Analysis of Metagenomic Data 2014

Faculty of Science

Statistical Analysis of Metagenomic Data

Viktor Jonsson

Licentiate thesis
Department of Mathematical Sciences
University of Gothenburg

UNIVERSITY OF GOTHENBURG

THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

Statistical analysis of metagenomic data

Viktor Jonsson

CHALMERS | GÖTEBORGS UNIVERSITET

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2014

Statistical analysis of metagenomic data
Viktor Jonsson
NO 2014:22
ISSN 1652-9715

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX.
Printed in Göteborg, Sweden 2014

Statistical analysis of metagenomic data

Viktor Jonsson

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg

Abstract

Metagenomics is the study of microbial communities on the genome level by direct sequencing of environmental and clinical samples. Recently developed DNA sequencing technologies have made metagenomics widely applicable and the field is growing rapidly. The statistical analysis is however challenging due to the high variability present in the data which stems from the underlying biological diversity and complexity of microbial communities. Metagenomic data is also high-dimensional and the number of replicates is typically few. Many standard methods are therefore unsuitable and there is a need for developing new statistical procedures.

This thesis contains two papers. In the first paper we perform an evaluation of statistical methods for comparative metagenomics. The ability to detect differentially abundant genes and control error rates is evaluated for eleven methods previously used in metagenomics. Resampled data from a large metagenomic data set is used to provide an unbiased basis for comparisons between methods. The number of replicates, the effect size and the gene abundance are all shown to have a large impact on the performance. The statistical characteristics of the evaluated methods can serve as a guide for the statistical analysis in future metagenomic studies. The second paper describes a new statistical method for the analysis of metagenomic data. The underlying model is formulated within the framework of a hierarchical Bayesian generalized linear model. A joint prior is placed on the variance parameters and shared between all genes. We evaluate the model and show that it improves the ability to detect differentially abundant genes.

This thesis underlines the importance of sound statistical analysis when the data is noisy and high-dimensional. It also demonstrates the potential of statistical modeling within metagenomics.

Keywords: Metagenomics, Statistical methods, Hierarchical Bayesian models, Statistical power, False discovery rate, Environmental genomics, Generalize linear models, Count data.

Acknowledgements

First of all I would like to thank my supervisor Erik Kristiansson. You are a source of endless inspiration and energy and your constant encouragement and new perspectives have had fundamental impact on me. I would also like to thank my co-supervisor Olle Nerman for his deep insights into statistics and ever so relevant comments.

To all my colleagues at the mathematics department, you make each day of work enjoyable. Unfortunately I cannot name you all but there are a still a few names that I simply cannot omit from this thesis. Fredrik Boulund, Mariana Pereira, Anna Larsson, Anna Johnning and Henrike Häbel, you all deserve special thanks.

To my family and friends, you make all this worthwhile. I hope you will see a little more of me from now on and that you can forgive me for my absence. :)

And finally, I have to give a special word of thanks to my mother whose daily question of “Hur är läget?” have kept me going for all these years.

Viktor Jonsson

Göteborg, November 2014

List of Papers

The licentiate thesis includes the following papers.

i. **Jonsson, V.**, Nerman, O., Kristiansson, E. (2014). A hierarchical bayesian model for ranking of genes in metagenomics based on differential abundance. *Manuscript*

ii. **Jonsson, V.**, Nerman, O., Kristiansson, E. (2014). Statistical evaluation of methods for comparative metagenomics. *Manuscript*

Below follows a list of relevant papers that are not included in this licentiate thesis.

- Viamontes, AE., Bengtsson-Palme, J., **Jonsson, V.**, Boulund, F., Rosvall, M., Kristiansson, E. (2014). The landscape of plasmid mediated antibiotic multiresistance. *Manuscript.*

Contents

Chapter 1

Introduction

The term metagenomics was first coined in 1998 and refers to the study of communities of microbes (Handelsman et al., 1998). As opposed to the classical genomics approach where individual organisms are studied one at a time, metagenomics enables sampling of a whole community simultaneously giving a snapshot view of all the genes present. This was originally performed using slow and expensive Sanger sequencing and the first metagenomes encompassed around 15000 DNA fragments (Sanger and Coulson, 1975; Healy et al., 1995). The introduction of next generation sequencing (NGS) methods, where millions of DNA fragments can be sequenced in parallel, has resulted in an increased throughput and lowered the price per base substantially (Schuster, 2008). This made metagenomics a much more widely applicable methodology and in a recent studies encompass several millions of fragments each (Fierer et al., 2012; Yatsunenko et al., 2012; Ward et al., 2013). In addition bacteria who cannot be cultured in the laboratory can be studied using metagenomics (Schloss and Handelsman, 2005). The value of the new knowledge produced by metagenomics is immense, examples include the linking of human diseases to gut microbiota (Karlsson et al., 2013), the detection of novel enzymes in unculturable bacteria(Hess et al., 2011) and characterization of communities that impact energy uptake in the gut (Turnbaugh et al., 2006).

A typical modern metagenomic analysis starts with extracting the microbial DNA from an environmental sample. The DNA is then sequenced using high-throughput sequencing yielding a vast number of short DNA fragments called reads. Depending on the technology used the length of reads vary from and average of 30 base pairs up to 400, with some recent technologies producing even longer reads (Metzker, 2010). There are several ways to process the reads further in terms of quality control and assembling them into longer sequences. Next, the reads are quantified by matching them against a reference database of choice. This process is called "binning" as reads with similar origin are binned together into groups. Depending on the target of interest, the reads are matched against databases containing species specific sequences, groups of genes with similar function or individual genes. The end result is a list of bins and their corresponding number of reads in each metagenomic sample. More extensive descriptions of the metagenomic analysis can be found in recent papers (Wooley et al., 2010; Hugenholtz and Tyson, 2008).

Given a list of counts, the next step is often to statistically compare metagenomes in order to detect differences (Tringe et al., 2005). Unfortunately the statistical analysis is complicated by high levels of both biological and technical variability (Wooley and Ye, 2009). The biological variability stems from the high diversity and complexity of microbial communities (Delmont et al., 2011). The bacterial species composition is for example known to vary considerably between samples (David et al., 2014). In addition, many bacterial species have plastic genomes and exhibit large variability in the gene content, even between individual members of a population (Kashtan et al., 2014). There is also a considerable variation in the presence of other organisms including viruses (Reyes et al., 2010). Technical sources of errors include the handling of samples and extraction of DNA which can introduce biases towards certain species (Morgan et al., 2010). The sequencing is also known to introduce errors as well as be biased with respect to the GC-content and repetitive contents (Benjamini and Speed, 2012). Furthermore the number of genes being investigated is typically very large (tens of thousands) while the number of replicates are low requiring methods that have the power to detect differences yet the specificity to avoid false positives. All these factors contribute to making statistical inference of metagenomic data complex.

In the broad perspective the area of statistics for metagenomics is still largely unexplored (Knight et al., 2012). Several statistical methods have been applied to metagenomics data but few novel ones have been developed (see *paper I*). The approaches tried include zero inflation (Paulson et al., 2013), non-parametric tests (Segata et al., 2011) and moderation of variance with the method presented in *paper II* appended to this thesis. Which of the many factors that are the most important to account for is yet unknown. In terms of statistical development a comparison can be made to the field of transcriptomics and the statistical analysis of microarrays. Microarrays enabled the analysis of several thousands of RNA transcripts simultaneously and were introduced in 1995 (Schulze and Downward, 2001). Initially many aspects of the variability were unknown and even whether the use of replicates was necessary (Lee et al., 2000). However, due to the low reproducibility the importance of proper normalization between samples and the benefits of new statistical methods soon became apparent. The example posed by microarrays underlines the potential of proper modeling the variability in large-scale molecular data. This suggests that further improvements can be made in the statistical analysis of metagenomic data and a similar development of dedicated statistical methods is needed to enable its full potential.

Chapter 2

Summary of papers

2.1 Paper I – Statistical evaluation of methods for comparative metagenomics

The aim of *Paper I* is to provide a comprehensive evaluation of statistical approaches for comparative metagenomics. The evaluation includes eleven methods that have been previously used on metagenomic data including both standard statistical tests as well as recently developed methods. The evaluation focuses on three different parts; the ability to rank genes based on differential abundance, the distribution of p-values under the null hypotheses as well as the ability to control the false discovery rate. In addition, the performance of the methods was measured for varying group sizes (number of replicates), effect sizes (fold change) and raw abundance (mean counts). The combination of these analyses is intended to give a full picture of the methods performance on metagenomic data both in terms of power to detect differences but at the same time controlling false positive rates.

Many comparative studies rely on a combination of data simulated from statistical models and real data sets. However, simulated datasets are based on model assumptions and inherently contain biases towards different models. Real data is free from these assumptions but there is no information about which genes are differentially abundant. In *paper I* we use resampled data from a real metagenomic data set to avoid the biases of simulated data but still have control of differential abundance. The resampled data is generated by randomly selecting a subsample to form an empirical null distribution. Differential abundance is then simulated by thinning the observed counts. The resampled data thus provides a realistic setting compared to simulated data.

The analysis showed that there were considerable differences between methods. This includes large differences in ranking performance between methods. Most methods also had biases in their p-value estimates with some methods being conservative and some optimistic. The most obvious aspect was the ability to handle overdispersion and methods that were unable to capture the variability in the data had a substantially lower performance in all aspects measured. Two methods were observed to have high performance across most settings. These were a generalized linear model based on an overdispersed Poisson distribution as well as the recently developed metagenomeSeq (Paulson et al., 2013). The study provides a guide to method selection for future analyses of metagenomic data.

2.2 Paper II – A hierarchical Bayesian method for the ranking of genes based on differential abundances

The focus of *paper II* is a new statistical method for detecting differentially abundant genes in metagenomic data. The model is formulated within the framework of a generalized linear model using a log link and the base variability of a Poisson distribution. The model captures the variability of the data by adding a random effect using a normal distribution with a gene specific overdispersion parameter. The key feature of the model is a global prior on the overdispersion parameters shared between all genes. This stabilizes the variance estimates of each gene. Similar methods that share variance information have proved very useful when analyzing microarrays (Smyth, 2004). However, the hierarchical structure of the model on top of the Poisson distribution makes it hard to treat analytically and MCMC (Markov Chain Monte Carlo) is used to fit the model to the data. The method is evaluated using both simulated data and data resampled from a real metagenome. By comparing to a nested model that does not share variance information we show that adding the global prior has a positive impact on the ability rank genes based on differential abundance. The conclusion is that modeling the variability of the data has large impact in metagenomics.

Chapter 3

Future work

As stated paper one aims to be a comprehensive evaluation of statistical methods for metagenomics. There are two key additions needed to fulfill this aim. Firstly there are a few methods for metagenomics that have not been included both recent developments and older methods. Most important among these are methods primarily developed for RNAseq. Even though these are not specifically developed for metagenomics they target overdispersed count data and have been used in a few metagenomic studies. Notable examples include edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010) and Voom(Law et al., 2014). The second necessary addition is the inclusion of another dataset to use for resampling. The dataset currently used is relevant but is limited to the human gut and the slightly outdated sequencing technique of 454 pyrosequencing (Yatsunenko et al., 2012). A second data set sequenced with Illumina technology will be added.

With regards to *paper II* the statistical method described has been shown to perform very well on the resampled metagenomic data. However the Markov Chain Monte Carlo implementation used is unable to handle large amounts of data. The current analyses in the paper were all made using JAGS (Just Another Gibbs Sampler) which has many merits (Plummer, 2003). It has both an easy way to formulate models and a good interface with R. In addition it features a module intended to improve the performance of generalized linear models. Even though this module does improve the convergence in out model it has a bug which causes the JAGS to crash when too many samples are taken. The results in the article have been generated using the glm-module however results have later been validated by running JAGS without the glm-module. Unfortunately sampling without the glm-module runs takes a considerably longer time. The problem can be solved by either using another sampler or by possibly re-parameterizing the model to improve mixing.

There are still several aspects of metagenomic data that are unexplored and could present interesting topics for research. For example recent studies have shown that metagenomic data may exhibit an increase in the number of zeros present due to sampling biases (Paulson et al., 2013). The generalized linear model framework used in paper II can be extended to handle zero-inflation. In a same way that paper II investigated whether a global prior on the variance would increase the ability to detect differentially abundant genes in metagenomics the benefits of zero-inflation could be evaluated in a future study.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72.

David, L. a., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, a. S., Varma, Y., Fischbach, M. a., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–63.

Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., Hirsch, P. R., and Vogel, T. M. (2011). Accessing the soil metagenome for studies of microbial diversity. *Applied and environmental microbiology*, 77(4):1315–24.

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., and Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52):21390–5.

Handelsman, J., Rondon, M., Brady, S., Clardy, J., and Goodman, R. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, 5.

Healy, F. G., Ray, R. M., Aldrich, H. C., Wilkie, A. C., Ingram, L. O., and Shanmugam, K. T. (1995). Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Applied Microbiology and Biotechnology*, 43(4):667–674.

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., Mackie, R. I., Pennacchio, L. a., Tringe, S. G., Visel, A., Woyke, T., Wang, Z., and Rubin, E. M. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–7.

Hugenholtz, P. and Tyson, G. W. (2008). Microbiology: metagenomics. *Nature*, 455(7212):481–3.

Karlsson, F., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C., Fagerberg, B., J, N., and F, B. (2013). Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498:99–103.

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., and Chisholm, S. W. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science*, 344(6182):416–20.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. a., Hugenholtz, P., van der Lelie, D., Meyer, F., Stevens, R., Bailey, M. J., Gordon, J. I., Kowalchuk, G. a., and Gilbert, J. a. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nature biotechnology*, 30(6):513–20.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*, 97(18):9834–9839.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46.

Morgan, J. L., Darling, A. E., and Eisen, J. a. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PloS one*, 5(4):e10209.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–2.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003).*

Reyes, A., Haynes, M., Hanson, N., Angly, F., Heath, A., Rohwer, F., and Gordon, J. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304):334–8.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40.

Sanger, F. and Coulson, A. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448.

Schloss, P. and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome biology*, 6(8):229.

Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays–a technology review. *Nature cell biology*, 3(8):E190–5.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16–8.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6):R60.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):–.

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. a., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P., and Rubin, E. M. (2005). Comparative metagenomics of microbial communities. *Science (New York, N.Y.)*, 308(5721):554–7.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. a., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31.

Ward, T. L., Hosid, S., Ioshikhes, I., and Altosaar, I. (2013). Human milk metagenome: a functional capacity analysis. *BMC microbiology*, 13(1):116.

Wooley, J., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667.

Wooley, J. and Ye, Y. (2009). Metagenomics: Facts and Artifacts, and Computational Challenges*. *Journal of computer science and technology*, 25(1):71–81.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. a., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–7.