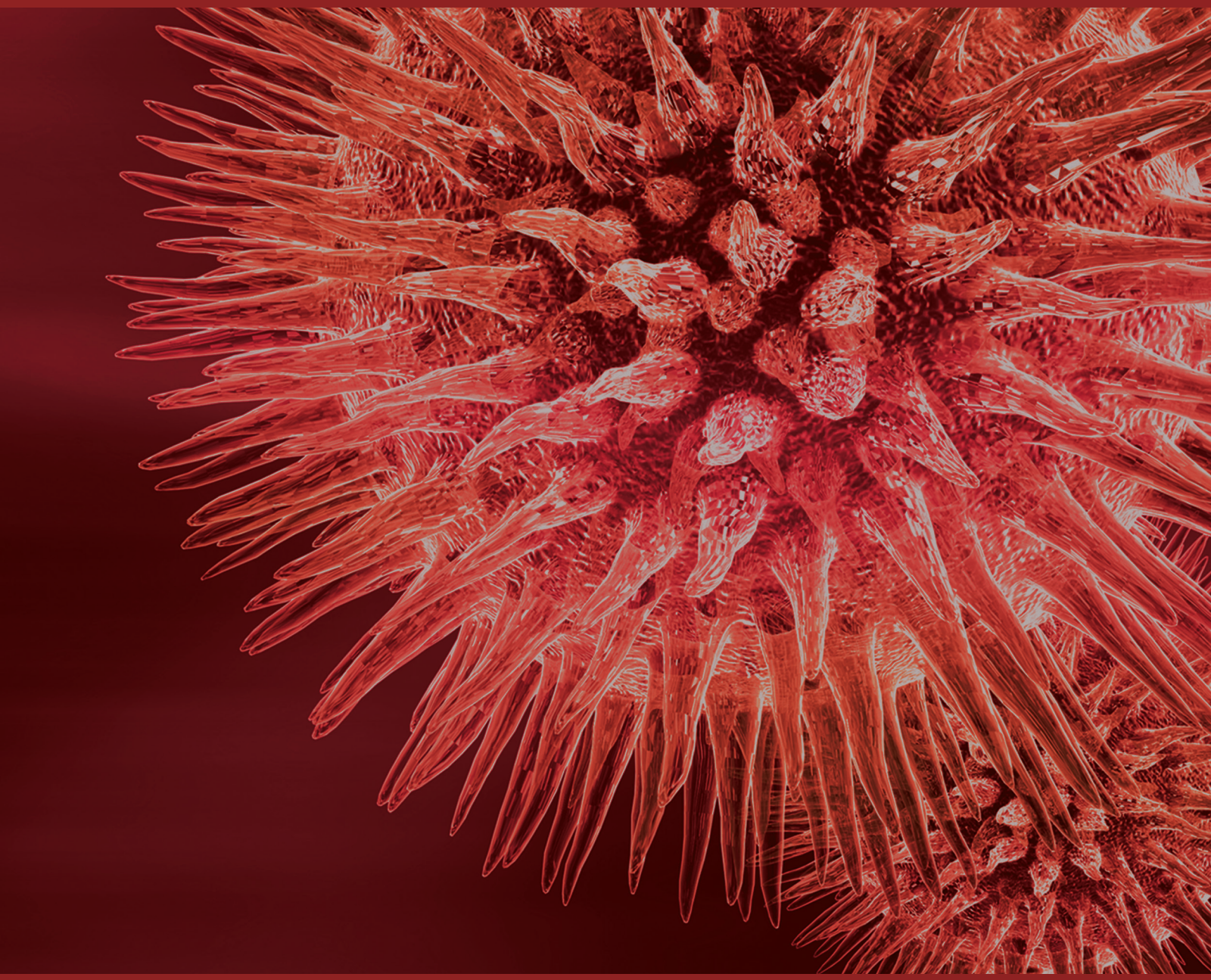
BioMed Research International

Bioinformatics Methods and Biological Interpretation for Next-Generation Sequencing Data

Guest Editors: Guohua Wang, Yunlong Liu, Dongxiao Zhu, Gunnar W. Klau, and Weixing Feng



Bioinformatics Methods and Biological Interpretation for Next-Generation Sequencing Data

BioMed Research International

Bioinformatics Methods and Biological Interpretation for Next-Generation Sequencing Data

Guest Editors: Guohua Wang, Yunlong Liu, Dongxiao Zhu, Gunnar W. Klau, and Weixing Feng

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 690873, 2 pages
http://dx.doi.org/10.1155/2015/690873

Editorial

Bioinformatics Methods and Biological Interpretation for Next-Generation Sequencing Data

Guohua Wang,[1] Yunlong Liu,[2] Dongxiao Zhu,[3] Gunnar W. Klau,[4] and Weixing Feng[5]

[1]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[2]*Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[3]*Department of Computer Science, Wayne State University, Detroit, MI 48202, USA*
[4]*Life Sciences, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, Netherlands*
[5]*Pattern Recognition and Intelligent System Institute, Automation College, Harbin Engineering University, Harbin, Heilongjiang 150001, China*

Correspondence should be addressed to Guohua Wang; ghwang@hit.edu.cn

Received 24 June 2015; Accepted 24 June 2015

Next-generation sequencing (NGS) technologies have revolutionarily reshaped the landscape of "-omics" research areas and their effects are becoming increasingly widespread. With its significantly lower costs and higher throughput, NGS has been applied to genome, transcriptome, and epigenome research. The plethora of information that emerges from large-scale next-generation sequencing experiments has triggered the development of bioinformatics tools and method for efficient analysis, interpretation, and visualization of NGS data. Such methods and tools will substantially promote the life-science community to better and efficiently help understand the underlying biological principles and mechanisms. This special issue mainly focuses on the original research articles as well as review articles that develop new bioinformatics approaches, present novel platforms and systems, and describe concise models well explaining the biological context and application in relation to genetics, metagenomics, and clinical study from NGS data.

This special issue contains nine papers. Two papers discuss the application of NGS data analysis in metagenomics and one paper presents R package for metagenomic systems biology analysis. One review paper discusses the software to detect alternative splicing isoforms from deep sequencing data. The other five papers are related to application of NGS data integration in genomics, genetics, and epigenetics.

In "mmnet: An R Package for Metagenomics Systems Biology Analysis," the authors developed R package, mmnet, to implement community-level metabolic network reconstruction and also implement a set of functions for automatic analysis pipeline construction. The package has substantial potentials in metagenomic studies that focus on identifying system-level variations of human microbiome associated with disease.

The paper "Constructing a Genome-Wide LD Map of Wild *A. gambiae* Using Next-Generation Sequencing" sequenced the genomes of nine individual wild *A. gambiae* mosquitoes using next-generation sequencing technologies. And 2,219,815 common single nucleotide polymorphisms (SNPs) were detected. Nearly one million SNPs that were genotyped with 99.6% confidence were extracted from these high-throughput sequencing data. Based on these SNP genotypes, the authors constructed a genome-wide linkage disequilibrium (LD) map for wild *A. gambiae* mosquitoes from malaria-endemic areas in Kenya and made it available through a public website.

The paper entitled "How to Isolate a Plant's Hypomethylome in One Shot" provided an easy, fast, and cost-effective tool to obtain a plant's hypomethylome (the nonmethylated part of the genome) by an optimized methyl filtration protocol with subsequent next-generation sequencing, in essence

a variant of MRE-seq. The hypomethylomes which were identified in three plant species, *Oryza sativa*, *Picea abies*, and *Crocus sativus*, showed clear enrichment in genes and their flanking regions. This method is extremely conducive to studying and understanding the genomes of nonmodel organisms.

In "Genetic Interactions Explain Variance in Cingulate Amyloid Burden: An AV-45 PET Genome-Wide Association and Interaction Study in the ADNI Cohort," the authors performed a genome-wide association study (GWAS) and a genome-wide interaction study (GWIS) of an amyloid imaging phenotype, using the data from Alzheimer's Disease (AD) Neuroimaging Initiative. The GWAS analysis revealed significant hits within or proximal to APOE, APOC1, and TOMM40 genes. The GWIS analysis yielded 8 novel SNP-SNP interaction findings that warrant replication and further investigation.

The paper "Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites" identified the functional transcription factor binding sites in gene regulatory region by integrating the DNase I hypersensitive sites with known position weight matrices. The authors present a model-based computational approach to predict a set of transcription factors that potentially cause such differential gene expression in cervical cancer HeLa S3 cell and HelaS3-ifna4h cell. This model demonstrated the potential to computationally identify the functional transcription factors in gene regulation.

The paper "Survey of Programs Used to Detect Alternative Splicing Isoforms from Deep Sequencing Data In Silico" is a review paper. Alternative splicing (AS) is very important for gene expression and protein diversity. First the authors summarized the alternative splicing forms and the means of selective splicing. Then the authors described the numerous methods for the read mapping of RNA-seq data and alternative types of splicing prediction software. At last, HMMSplicer, SOAPsplice, TopHat, and STAR were used to evaluate the performance of alternative splicing isoforms detection.

The article "MicroRNA Promoter Identification in *Arabidopsis* Using Multiple Histone Marker" was devoted to a computational strategy, which identified the promoter regions of most microRNA genes in *Arabidopsis*, using the genome wide profiles of nine histone markers. Based upon the assumption that the distributions of histone markers around the transcription start sites (TSSs) of microRNA genes are similar with the TSSs of protein coding gene, the Support Vector Machine (SVM) was used to identify 42 independent miRNA TSSs and 132 miRNA TSSs which are located in the promoters of upstream genes. The annotation of microRNA TSSs will provide the measurements regarding the initiation of transcription and better understanding of microRNA regulation.

The paper "454-Pyrosequencing Analysis of Bacterial Communities from Autotrophic Nitrogen Removal Bioreactors Utilizing Universal Primers: Effect of Annealing Temperature" carried out a metagenomic analysis (pyrosequencing) of total bacterial diversity including Anammox population in five autotrophic nitrogen removal technologies, two bench-scale (MBR and low temperature CANON) and three full-scale (Anammox, CANON, and DEMON), by optimization of primer selection and PCR conditions. The pyrosequencing data showed that annealing temperature of 45°C yielded the best results in terms of species richness and diversity for all bioreactors analyzed.

The paper entitled "Active Microbial Communities Inhabit Sulphate-Methane Interphase in Deep Bedrock Fracture Fluids in Olkiluoto, Finland" investigated active microbial communities of deep crystalline bedrock fracture water from seven different boreholes in Olkiluoto (Western Finland), using bacterial and archaeal 16S rRNA, *dsr*B, and *mcr*A gene transcript targeted 454 pyrosequencing. The results demonstrated that active and highly diverse but sparse and stratified microbial communities inhabited the Fennoscandian deep bedrock ecosystems.

Acknowledgments

The guest editors heartily thank all authors for their excellent contributions and patience in undertaking revisions of their manuscripts. We would like to acknowledge the numerous reviewers for their professional effort that helped to improve the quality of the selected articles in this special issue. We hope that the readers will find interesting NGS methods and application in the issue.

Guohua Wang
Yunlong Liu
Dongxiao Zhu
Gunnar W. Klau
Weixing Feng

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 861402, 10 pages
http://dx.doi.org/10.1155/2015/861402

Research Article

MicroRNA Promoter Identification in *Arabidopsis* Using Multiple Histone Markers

Yuming Zhao,[1,2] Fang Wang,[3] and Liran Juan[3]

[1]*State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin, Heilongjiang 150001, China*
[2]*Information and Computer Engineering College, Northeast Forestry University, Harbin, Heilongjiang 150001, China*
[3]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*

Correspondence should be addressed to Yuming Zhao; zymyoyo@hotmail.com

Received 11 December 2014; Accepted 12 March 2015

Academic Editor: Cheol Yong Choi

A microRNA is a small noncoding RNA molecule, which functions in RNA silencing and posttranscriptional regulation of gene expression. To understand the mechanism of the activation of microRNA genes, the location of promoter regions driving their expression is required to be annotated precisely. Only a fraction of microRNA genes have confirmed transcription start sites (TSSs), which hinders our understanding of the transcription factor binding events. With the development of the next generation sequencing technology, the chromatin states can be inferred precisely by virtue of a combination of specific histone modifications. Using the genome-wide profiles of nine histone markers including H3K4me2, H3K4me3, H3K9Ac, H3K9me2, H3K18Ac, H3K27me1, H3K27me3, H3K36me2, and H3K36me3, we developed a computational strategy to identify the promoter regions of most microRNA genes in *Arabidopsis*, based upon the assumption that the distribution of histone markers around the TSSs of microRNA genes is similar to the TSSs of protein coding genes. Among 298 miRNA genes, our model identified 42 independent miRNA TSSs and 132 miRNA TSSs, which are located in the promoters of upstream genes. The identification of promoters will provide better understanding of microRNA regulation and can play an important role in the study of diseases at genetic level.

1. Introduction

MicroRNAs (miRNAs) are small (~22 nucleotides) noncoding RNAs, which are processed to ~70-nucleotide precursors and subsequently to the mature form by endonucleases [1, 2]. miRNAs are disseminated throughout the genome. They can be found in intergenic regions, intronic regions of protein coding genes, or intronic and exonic regions of noncoding RNAs. They have many regulatory functions in complex organisms. It is known that a single miRNA can influence the expression of thousands of genes, thus controlling one-third of the human genome [3]. Recent studies have showed their association with human diseases and cancer [4, 5] and indicate that miRNA expression, whether intronic or intergenic, may be complex and varied among tissues, cell types, and disease states [6–8].

The promoter of a gene is a crucial control region for its transcription initiation [9, 10]. To make out the mechanism of the activation of miRNA genes, it is required to locate their core promoter regions. It has been noticed that promoter regions contain characteristic features that can be used to distinguish them from other parts of the genome. These features may be grouped into three types: signal, context, and structure features [11]. Signal features are biologically functional regions including core-promoter elements [11, 12], some short modular transcription factor binding sites (TFBSs), and CpG-islands [13], which play important roles in assembly of transcriptional machinery. Context features are extracted from the genomic content of promoters as a set of n-mers (n-base-long nucleotide sequences) whose statistics are estimated from training samples [14]. Structure features are derived from DNA three-dimensional structures, which play important roles in guiding DNA-binding proteins to target sites efficiently [15, 16].

However, only a small portion of the human miRNAs has confirmed transcription start sites (TSSs). Imperfect

knowledge of the start sites of primary miRNA transcripts has limited our ability to study the promoter sequence features and further identify the transcription factor binding events. All existing promoter prediction methods for protein coding genes may not be suitable for miRNA genes, since they were not built based on the core promoters of miRNA genes. Hence, some studies have predicted the human pri-miRNAs boundary and regulatory region by EST [17, 18], sequence feature [19], and evolutionarily conservation [20]. Recently, several studies that utilized high-throughput genomic techniques identified the likely location of human miRNA TSSs [7, 8, 21–25]. The sequencing of 5′ transcript ends [26] and genome tilling microarrays (ChIP-chip) for RNA polymerase II [27] have been used to identify proximal promoters of miRNAs in *Arabidopsis*. Although numerous prediction models were developed for identifying miRNA promoters or TSSs, inadequate evidence was revealed to elucidate relationships between miRNA genes and transcription factors (TFs) due to lack of experimental validation.

With the next generation sequencing technology development, the genome-wide chromatin profiles have been detected. Using the combinations of specific histone modifications, chromatin states correlate with regulator binding, transcriptional initiation, and elongation; enhancer activity and repression can be inferred more precisely. Using biologically meaningful and spatially coherent combinations of chromatin marks, two studies have proposed a novel approach for discovering chromatin states, in a systematic de novo way across the whole genome based on a multivariate hidden Markov model (HMM) [28, 29] which explicitly modeled mark combinations. The chromatin marks included histone acetylation marks, histone methylation marks, and CTCF/Pol2/H2AZ. By analyzing the genome-wide occupancy data for these chromatin marks, the chromatin states were definitely classified. Even though states were learned de novo based solely on the patterns of chromatin marks and their spatial relationships, they correlated strongly with upstream and downstream promoters, 5′-proximal and distal transcribed regions, active intergenic regions, and repressed and repetitive regions. And these chromatin states were distinguished into six broad classes including promoter states, enhancer states, insulator states, transcribed states, repressed states, and inactive states according to the present/absent condition of the combination of chromatin marks.

Recently, genome-wide maps of nine histone modifications produced by ChIP-Seq were used to describe the chromatin patterns in *Arabidopsis* [30]. Previous study has found that miRNA and protein coding genes share similar mechanisms of regulation by chromatin modifications [31]. Based upon the assumption that the distributions of histone markers around the TSSs of miRNA genes are similar to the TSSs of protein coding genes, we have developed a computational strategy to identify the promoter regions of most miRNA genes. Comparing to HMM, the Support Vector Machine (SVM) has better performance in binary classification. In this study, SVM was used to distinguish the distributions of 9 histone markers in promoter regions and nonpromoter regions.

2. Method

2.1. Data Description. In the previous study, the ChIP-Seq experiments of nine histone modifications, H3K4me2, H3K4me3, H3K9Ac, H3K9me2, H3K18Ac, H3K27me1, H3K27me3, H3K36me2, and H3K36me3, were produced in the aerial tissue of 2-week-old *Arabidopsis* plants [30]. The whole genome profiles of nine ChIP-Seq experiments were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (accession number GSE28398). 35 bps color-space reads for each of the histone markers were aligned to TAIR 8 *Arabidopsis thaliana* reference genome. Here, we converted the genome position of each read from TAIR 8 genome assembly to TAIR 10 genome assembly.

The gene annotation of TAIR 10 genome assembly was downloaded from *Arabidopsis* Information Resource (TAIR) (https://www.arabidopsis.org/), in which more than 27,000 protein coding genes were annotated. *Arabidopsis* miRNAs annotation was downloaded from miRNA registry [32] (miRBase, v20), and 298 miRNAs were annotated in miRBase.

2.2. ChIP-Seq Data Processing. In order to retrieve the histone modifications patterns around the transcription start sites of protein coding gene, we first divided the genomic regions neighboring TSS into 100 bp bins. We wanted to compare the same regions in the genome for the number of reads they have in nine different histone modification libraries, so we absolutely normalized the raw data to reads per million per bin (RPM). We calculated the number of reads that fall into an individual bin divided by the number of reads in the sample data set and then multiplied by 10^6 to get the value per million. In this way we got an RPM track of 100 base bins covering the genome; thus samples with different numbers of reads become comparable. The RPM formula is as follows:

$$\mathrm{RPM}_i = R_i * \frac{10^6}{N}. \tag{1}$$

Here, R_i represents the number of reads falling into the ith bin and N represents the total number of mapped reads.

The histone binding pattern nearby the TSSs of all protein coding genes is retrieved as positive set, and the ones of 10,000 random positions are retrieved as negative set.

2.3. Support Vector Machine. In this study, the Support Vector Machine (SVM) was used to classify the TSSs and random regions based on the profiles of nine histone markers. The SVM is described as follows.

Given a training data D, a set of n points of the form

$$D = \left\{(x_i, y_i) \mid x_i \in R, y_i \in \{-1, 1\}\right\}_{i=1}^{n}, \tag{2}$$

where x_i is the reads count in each bin around the ith gene's TSS and the y_i is either 1 or −1, indicating the two classes to be classified as the real transcription start sites versus random genomic regions.

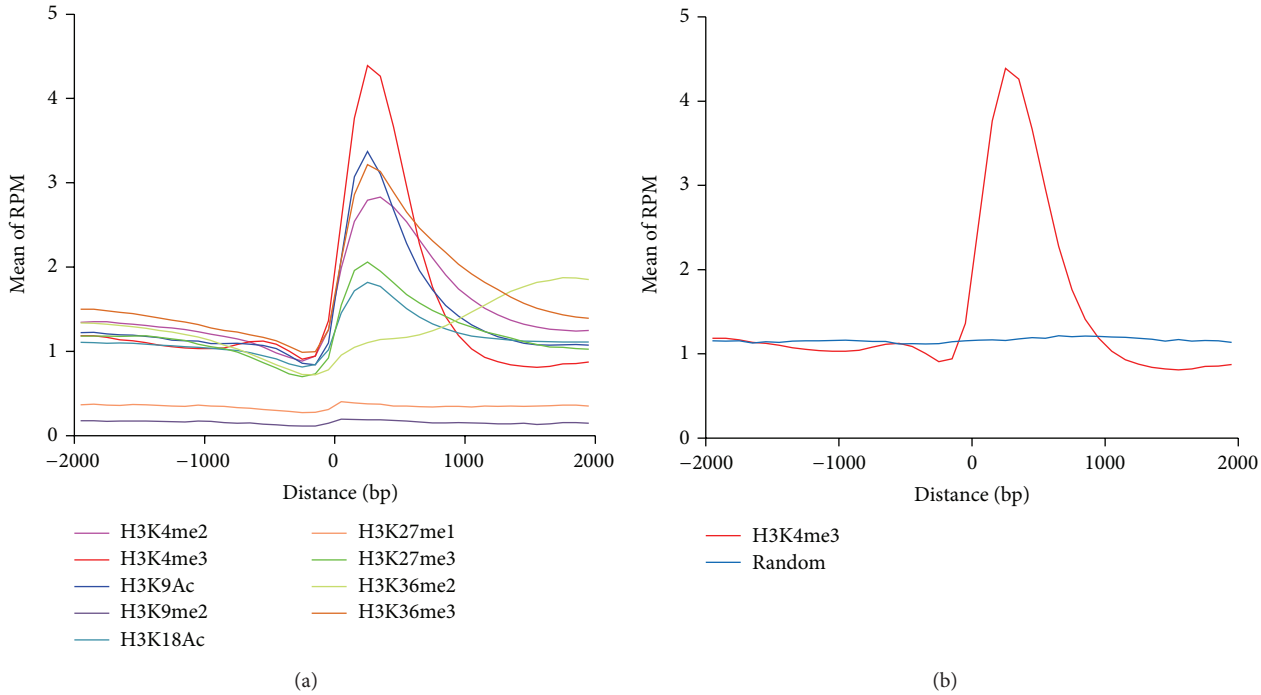


Figure 1: The distribution of histone markers around TSS of protein coding genes. (a) The ChIP-Seq-derived histone modifications patterns around the TSS of protein coding gene in *Arabidopsis*. The RPM distributions of nine histone markers including H3K4me2, H3K4me3, H3K9Ac, H3K9me2, H3K18Ac, H3K27me1, H3K27me3, H3K36me2, and H3K36me3 were marked by different colors. (b) The pattern of ChIP-Seq-derived H3K4me3 around the TSS of protein coding gene (red curve) and random genomic region (blue) in *Arabidopsis*.

The decision function of SVM is

$$\operatorname{sgn}\left(\sum_{i=1}^{n} y_i a_i K(x_i, y_i) + \rho\right). \tag{3}$$

Here, $K(x_i, y_i)$ is the radial basis function (RBF) kernel, because of its good general performance and a few number of parameters (only two: C and γ).

We used the R package "e1017," which offers an interface to package LibSVM (version 2.6). To obtain SVM classifier with optimal performance, the penalty parameter C and the RBF kernel parameter γ are tuned based on the training set using the grid search strategy in e1017.

3. Result

3.1. The Histone Marker Distribution around TSS of Protein Coding Genes. The goal of this study was to use ChIP-Seq-derived histone marker data to identify transcription start sites of miRNAs in *Arabidopsis*. We first examined the 9-histone-marker pattern around the TSS of protein coding genes. We divided the genomic regions into multiple 100 bp bins and calculated the reads per million per bin (RPM) of each histone marker's fragments located in each of the bins within 2,000 bp upstream and downstream the TSS. Not surprisingly, most of these nine histone markers are enriched around the transcription start sites of protein coding genes and a peak of RPM can be found near the TSS (Figure 1(a)). While, among these nine histone markers, the H3K4me3 has the most significant peak, H3K9me2 and H3K27me1 have no peak in TSS. We also randomly selected 10000 genomic positions to examine the histone marker pattern. No such enrichment was found for H3K4me3 signal (Figure 1(b)) and other histone markers in random genomic regions. It suggests that the histone markers are strongly correlated with TSS. We can predict the promoter by examining the distribution of histone markers around TSS.

3.2. The Training and Prediction of SVM Using Nine Histone Markers. In this study, we used Support Vector Machine (SVM) to predict the TSSs of miRNA based on the profiles of 9 histone markers. We selected the fragment distribution derived from ChIP-Seq data of 9 histone markers around TSS in 27,000 protein coding genes as positive set and those on 10,000 random positions as negative set. To estimate the accuracy of our method, we used random half of the positive set and half of the negative set to train SVM and then predicted the remaining positive and negative set. The prediction probability was presented using characteristic curve (ROC curve), in which the abscissa is specificity that represents the false positive rate and the ordinate is sensitivity that represents the true positive rate. If the area under ROC curve (AUC) is bigger, the accuracy of prediction is higher. At first, we used the distribution pattern of one single histone marker to train SVM and nine ROC curves of predicted result were shown in Figures 2(a)–2(i). For each histone marker, 4 different histone patterns were picked up around TSS, which were 20, 15, 10, and 5 bins up and down TSS. Notably,

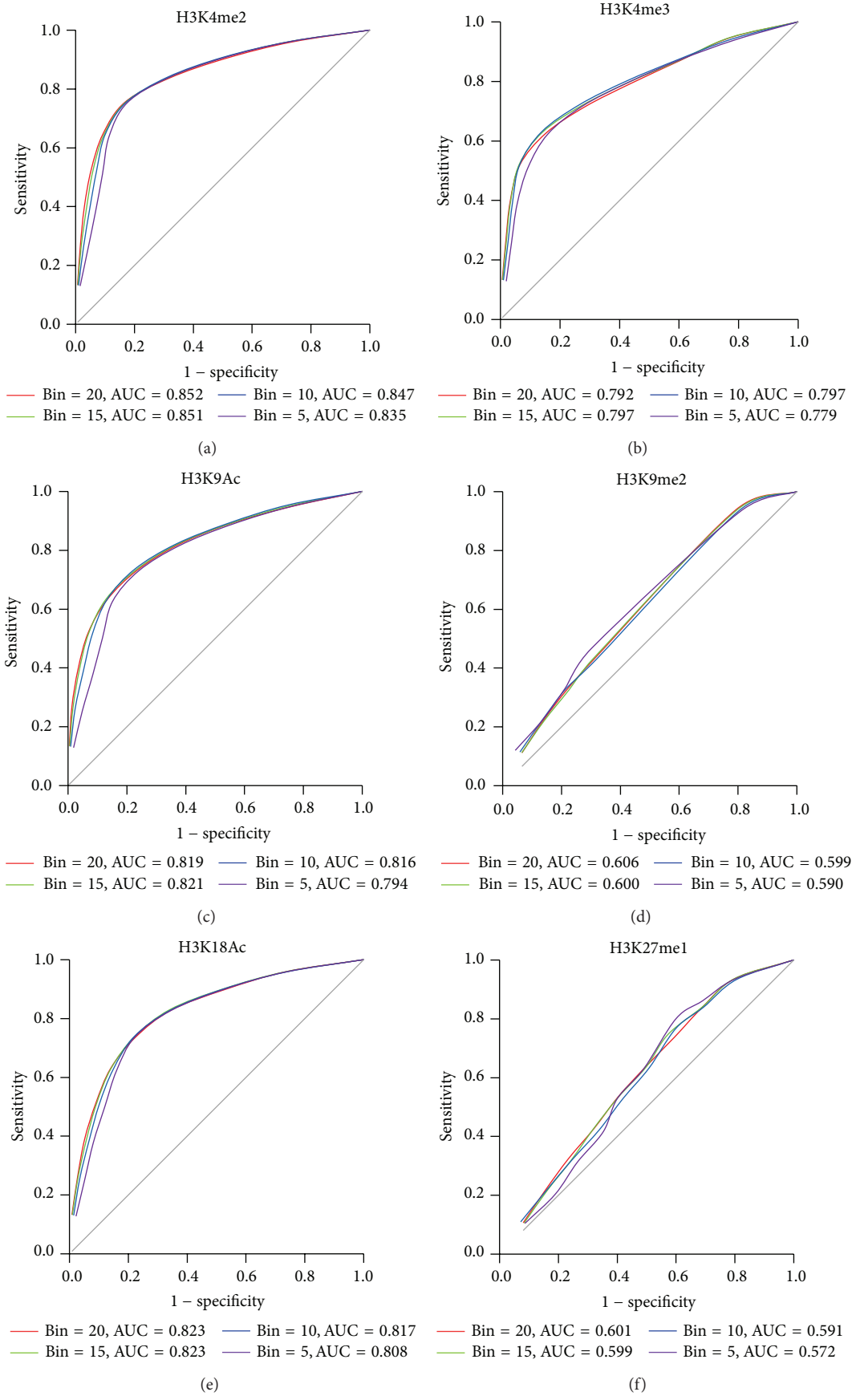


Figure 2: Continued.

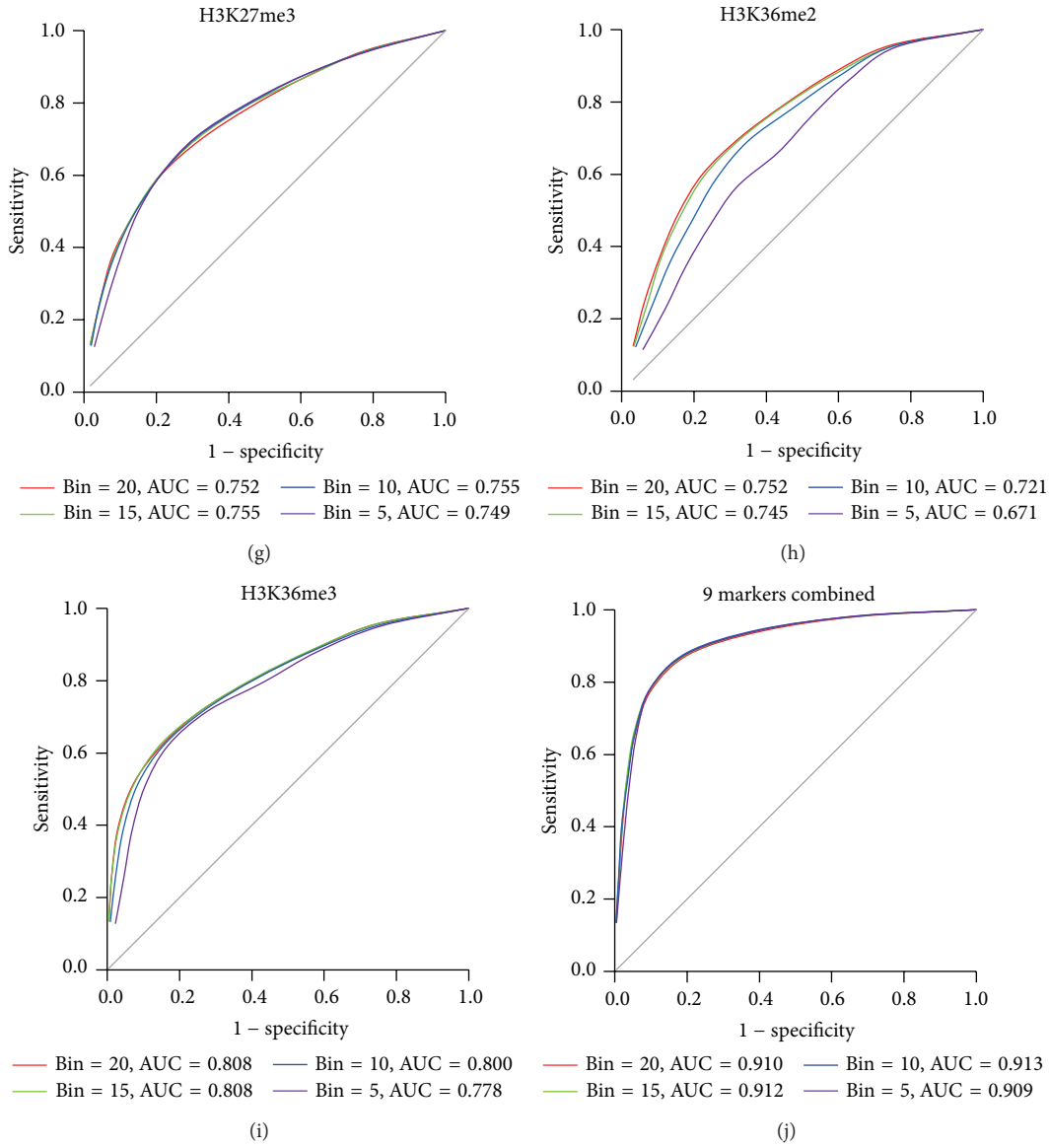


Figure 2: ROC curve for TSS prediction of protein coding genes with different histone markers. From (a) to (i), the ROC curve shows the sensitivity and specificity of the TSS prediction for protein coding genes with different histone marker. For each histone marker, the ROC curve was calculated within four different ranges around the TSS. For example, the red curve represents the ROC curve calculated within 20 bins up and 20 bins down of the TSS. The area under the curve (AUC) for each range around the TSS is shown in each graph. (j) The ROC curve for TSS prediction of protein coding genes with the combined nine histone markers.

the H3K4me3 has the biggest AUC (~0.85) and the prediction accuracy in each pattern selection based on different bin number is very similar. On the contrary, the H3K9me2 and H3K27me1 have the smallest AUC (~0.6). This result is very consistent with the enrichment of histone marker pattern around the TSS. In most histone marker predictions, the AUC scores are increased from 5 bins to 20 bins, which means the statistical power is increased. To get more accurate prediction, we combined all the nine histone markers to train SVM and predict TSS (Figure 2(j)). All the AUC scores of 4 different histone patterns based on bin number are above 0.9. The highest AUC score is 0.913 based on prediction of 10 bins. In the next step, we will predict the miRNA transcription start sites by integrating 9 histone markers around 10 bins of upstream and downstream TSSs.

3.3. The Prediction of TSSs of miRNA Genes. The objective of this study was to identify the TSSs of miRNAs by searching for histone marker patterns similar to those seen in the upstream regions of protein coding genes. The *Arabidopsis* genome has a large density of gene distribution, which is about one gene every 4-5 kb. The distance between miRNA and the TSS of its nearest upstream gene was calculated (Figure 3(a)), which showed miRNAs whose corresponding distance is <1 k, 1-2 k, or 2-3 k had the highest frequencies. 217 of 298 distances

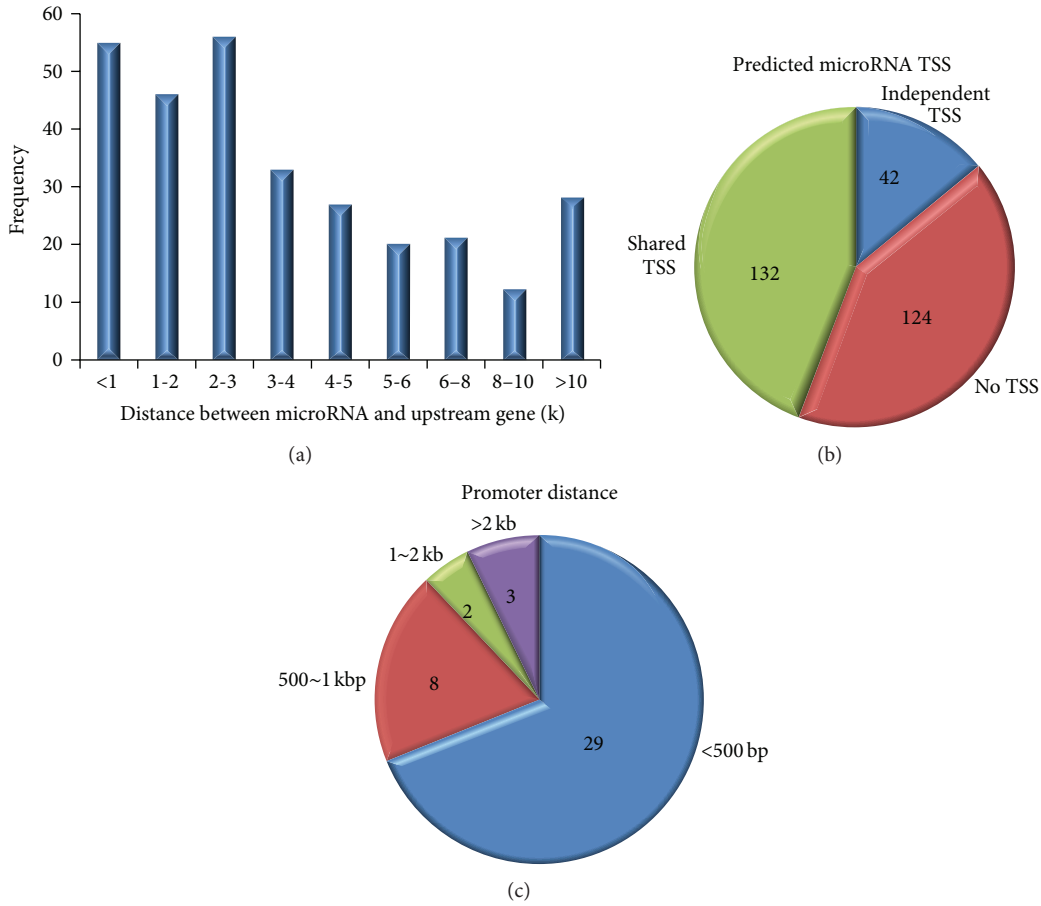


Figure 3: Features of predicted miRNA TSSs. (a) Histogram illustrating the distance between 298 miRNAs and their upstream genes. (b) The number of identified miRNA TSSs. The blue sector represents the 42 independent miRNA TSSs. The green sector represents the 132 predicted miRNA TSSs in the same position as the TSS of their upstream genes. The red sector represents 124 miRNAs that have no predicted TSS. (c) The distances between the predicted independent miRNA TSSs and their corresponding miRNAs.

between miRNAs and upstream TSSs are less than 5 k, and no miRNAs are far away from the nearest upstream TSS to 10 K. In this study, we focused our study on 298 miRNAs obtained from miRBase miRNA sequence database (version 20). For each miRNA, SVM was used to search the TSS of the primary miRNA up to 10 kbp upstream the mature miRNAs. We combined the profile of 9 histone markers in up and down 10 bins around TSS to train the SVM and then predict promoters of 298 miRNA genes. Using FDR $\leq$ 0.1, we identified 42 miRNAs which have independent TSSs (Table 1). Among 298 miRNAs, the predicted TSSs of 124 miRNAs were at the same position as the promoter of upstream nearest gene (Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/861402), and the remaining 132 miRNAs were not predicted (Figure 3(b)). We also calculated the distance between 42 independent TSSs and the corresponding miRNAs, which were shown in Figure 3(c). Obviously, most of the independent predicted promoters were much close to the corresponding miRNAs.

3.4. The Histone Pattern around Predicted miRNA TSSs. miRNAs ath-MIR167b and ath-MIR773b are selected to show the histone pattern around the predicted TSSs (Figure 4). miRNA ath-MIR167b and upstream gene AT4G19390 are head-to-head gene pair (Figure 4(a)). Our method identified an independent TSS of ath-MIR167b, which is 2 k far away from gene AT4G19390. The histone marker profiles showed two different peaks of combined nine histone markers, which suggests these two opposite genes had differently regulatory regions. miRNA ath-MIR773b and upstream gene AT1G35470 are in the same strand (Figure 4(b)). The predicted TSS of ath-MIR773b overlaps with the AT1G35470 promoter region. Only one histone marker peak can be found in the promoter region of AT1G35470, and no histone marker is enriched between ath-MIR773b and AT1G35470. One biological mechanism is that the miRNAs are transcribed with the upstream genes at the same time. We found that 25 out of 124 miRNAs with the same TSSs as upstream genes were intronic miRNAs (Supplementary Table 1), which means these miRNAs had the same regulatory region as host genes.

3.5. The Comparison with Other miRNA Promoter Identification Methods. In previous studies, 5′ rapid amplification

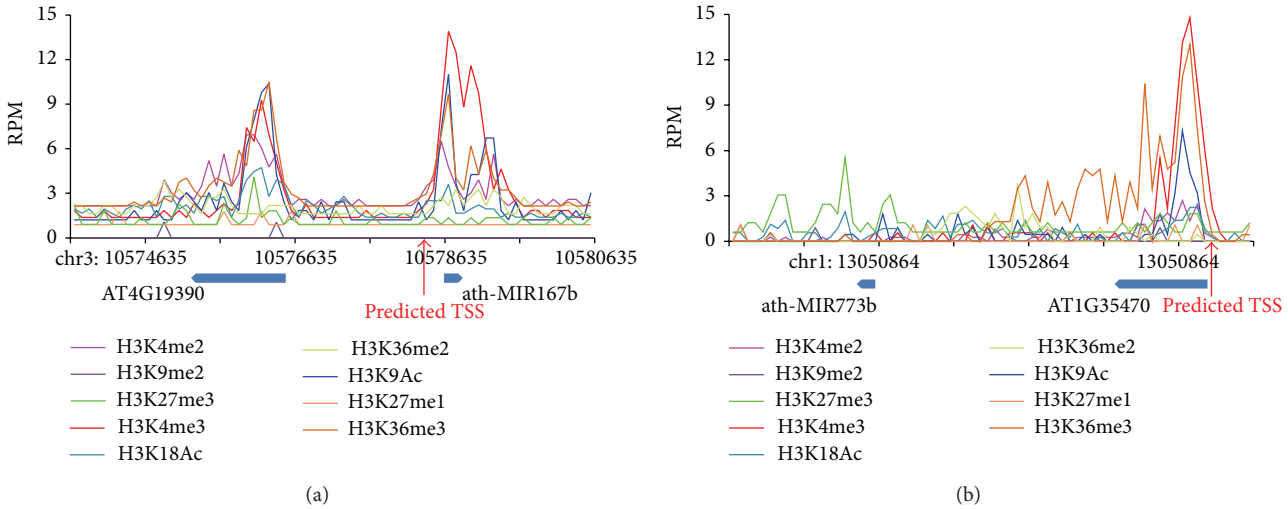


FIGURE 4: The histone pattern between predicted miRNA TSSs and upstream gene TSSs. (a) The example of independent predicted miRNA TSS. The first peak represents the position of the upstream gene TSS and the second peak represents the position of predicted miRNA TSS. (b) The example of predicted miRNA TSS has the same position as the TSS of upstream gene. Different histone markers are presented in different color.

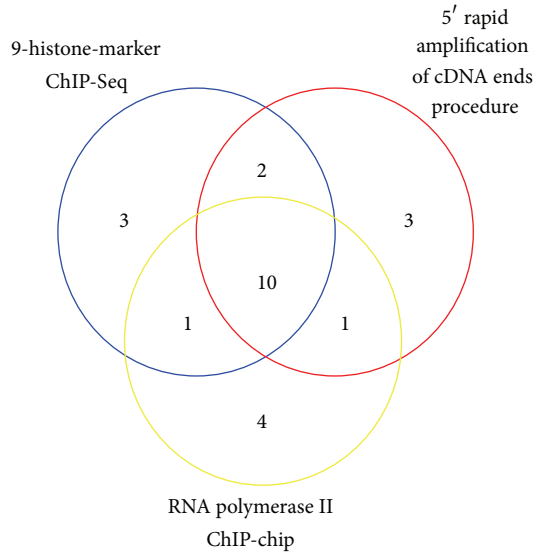


FIGURE 5: The overlapping of 16 microRNA TSSs identified by all three methods.

of cDNA ends procedure [26] and RNA polymerase II ChIP-chip experiment [27] have been used to determine promoters of miRNAs in *Arabidopsis*. Our study predicted 42 independent miRNA TSSs by 9 histone markers. 16 among 42 TSSs of miRNAs were also identified by other two methods. If a TSS position recognized by one method locates within 100 bp from TSS of the same miRNA identified by another method, this TSS was considered to be the same by these two methods. As we can see in Figure 5, 10 out of 16 miRNA TSSs were consistent for all three methods. One miRNA TSS was identified as being the same by our method and polymerase II ChIP-chip, but not by 5′ rapid amplification of cDNA ends procedure. Two miRNA TSSs were the same for our result and 5′ rapid amplification of cDNA ends procedure only.

4. Discussion

Annotation of the primary transcripts of miRNAs is extremely important to our understanding of the biological process of miRNAs and their regulatory targets. Although much progress has been made in promoter recognition, we are still far away from the goal of miRNA promoter identification. High-throughput DNA sequencing is rapidly changing the landscape of genomic research [33]. Recent studies using ChIP-Seq technology have revealed genome-wide transcription factor binding sites [34–36], RPol II binding sites and patterns associated with active transcription of coding genes [37, 38], and the distribution of histone modifications across the genome [38]. The modifications of the histones are found to be associated with transcription initiation and elongation [39], which made plenty of promoter prediction studies regarding histone modification as significant features. For example, H3K4me3 is enriched in the promoter regions, and H3K36me3 occurs at nucleosomes covering primary transcripts of actively expressed genes [40].

In this study, 9 histone markers including H3K4me2, H3K4me3, H3K9Ac, H3K9me2, H3K18Ac, H3K27me1, H3K27me3, H3K36me2, and H3K36me3 were used to predict miRNA promoters. Based upon the assumption that the distributions of histone markers around the TSSs of miRNAs are similar to the ones of protein coding genes, we developed a computational strategy to identify the promoter regions of all miRNA genes in *Arabidopsis*. Integrating 9

Table 1: 42 independent predicted miRNA transcription start sites.

Index	miRNA ID	miRNA name	Genome coordinates	TSS
1	MI0000989	ath-MIR171b	chr1:3961348–3961464(−)	3961764–3961864
2	MI0005386	ath-MIR830	chr1:4820355–4820549(−)	4820549–4820649
3	MI0000218	ath-MIR159b	chr1:6220646–6220841(+)	6220446–6220546
4	MI0001005	ath-MIR394a	chr1:7058194–7058310(+)	7055994–7056094
5	MI0019201	ath-MIR5630a	chr1:12011152–12011223(−)	12011523–12011623
6	MI0019211	ath-MIR5630b	chr1:12023526–12023597(−)	12023997–12024097
7	MI0000193	ath-MIR161	chr1:17825685–17825857(+)	17825485–17825585
8	MI0019208	ath-MIR5636	chr1:18549959–18550036(+)	18549659–18549759
9	MI0001078	ath-MIR406	chr1:19430078–19430277(−)	19431177–19431277
10	MI0019235	ath-MIR5652	chr1:23412989–23413436(−)	23413636–23413736
11	MI0000196	ath-MIR163	chr1:24884066–24884396(+)	24883966–24884066
12	MI0001425	ath-MIR414	chr1:25137456–25137563(−)	25137763–25137863
13	MI0000189	ath-MIR159a	chr1:27713233–27713416(−)	27713616–27713716
14	MI0015817	ath-MIR4228	chr1:28889375–28889532(+)	28889175–28889275
15	MI0005105	ath-MIR775	chr1:29422452–29422574(+)	29422052–29422152
16	MI0001013	ath-MIR396a	chr2:4142323–4142473(−)	4142673–4142773
17	MI0005109	ath-MIR779	chr2:9560761–9560923(+)	9560161–9560261
18	MI0020189	ath-MIR5995b	chr2:10026910–10027050(+)	10026310–10026410
19	MI0020188	ath-MIR5595a	chr2:10026910–10027050(−)	10027050–10027150
20	MI0000178	ath-MIR156a	chr2:10676451–10676573(−)	10676673–10676773
21	MI0000215	ath-MIR172a	chr2:11942914–11943015(−)	11943215–11943315
22	MI0017889	ath-MIR5021	chr2:11974711–11974881(−)	11975181–11975281
23	MI0000201	ath-MIR166a	chr2:19176108–19176277(+)	19176008–19176108
24	MI0001072	ath-MIR403	chr2:19415052–19415186(+)	19414952–19415052
25	MI0000208	ath-MIR167a	chr3:8108072–8108209(+)	8107972–8108072
26	MI0005383	ath-MIR827	chr3:22122760–22122936(−)	22123036–22123136
27	MI0000202	ath-MIR166b	chr3:22922206–22922325(+)	22921906–22922006
28	MI0002407	ath-MIR447a	chr4:1528134–1528370(−)	1529270–1529370
29	MI0017896	ath-MIR5026	chr4:7844496–7844688(+)	7842896–7842996
30	MI0005405	ath-MIR850	chr4:7845707–7845927(+)	7842907–7843007
31	MI0005440	ath-MIR863	chr4:7846597–7846899(+)	7842897–7842997
32	MI0015815	ath-MIR4221	chr4:8460516–8460662(+)	8459516–8459616
33	MI0000210	ath-MIR168a	chr4:10578635–10578772(+)	10578335–10578435
34	MI0000180	ath-MIR156c	chr4:15415418–15415521(−)	15415821–15415921
35	MI0019242	ath-MIR5658	chr4:18485438–18485531(−)	18486431–18486531
36	MI0000198	ath-MIR164b	chr5:287584–287736(+)	287484–287584
37	MI0000216	ath-MIR172b	chr5:1188207–1188301(−)	1188501–1188601
38	MI0000195	ath-MIR162b	chr5:7740598–7740708(−)	7740908–7741008
39	MI0019216	ath-MIR5643a	chr5:11667797–11667879(+)	11667197–11667297
40	MI0001014	ath-MIR396b	chr5:13611798–13611932(+)	13611698–13611798
41	MI0000211	ath-MIR168b	chr5:18358788–18358911(−)	18359011–18359111
42	MI0001075	ath-MIR405b	chr5:20632514–20632637(+)	20630514–20630614

histone markers profiles, the model based on SVM classifier identified 42 independent miRNA TSSs from total 298 miRNA genes, and 132 predicted miRNA TSSs were identified in the same position as the TSS of their upstream genes. We also found that 25 out of 124 miRNAs were intronic miRNAs, which suggest that most of the intronic miRNAs share promoter regions with their host genes. For the remaining genes, we lack the evidence to explain whether they share the promoter with the upstream genes or have independent promoter. So the identification of miRNA promoters in other tissues of *Arabidopsis*, in addition to the aerial tissue, will improve the annotation of primary transcripts of miRNAs.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contribution

Yuming Zhao designed the project. Fang Wang and Liran Juan performed the experiments and wrote the paper. Yuming Zhao revised the paper. All the authors read and approved the final manuscript.

Acknowledgments

The work was supported by the Fundamental Research Funds for the Central Universities (DL10BB02), State Key Laboratory of Tree Genetics and Breeding (Northeast Forestry University) (201207), and the Natural Science Foundation of China (61371179).

References

[1] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[2] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nature Reviews Genetics*, vol. 5, no. 7, pp. 522–531, 2004.

[3] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[4] T. Thum, P. Galuppo, C. Wolf et al., "MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure," *Circulation*, vol. 116, no. 3, pp. 258–267, 2007.

[5] E. G. Nikitina, L. N. Urazova, and V. N. Stegny, "MicroRNAs and human cancer," *Experimental Oncology*, vol. 34, no. 1, pp. 2–8, 2012.

[6] A. M. Monteys, R. M. Spengler, J. Wan et al., "Structure and activity of putative intronic miRNA promoters," *RNA*, vol. 16, no. 3, pp. 495–505, 2010.

[7] F. Ozsolak, L. L. Poling, Z. Wang et al., "Chromatin structure analyses identify miRNA promoters," *Genes & Development*, vol. 22, no. 22, pp. 3172–3183, 2008.

[8] X. Wang, Z. Xuan, X. Zhao, Y. Li, and M. Q. Zhang, "High-resolution human core-promoter prediction with CoreBoost-HM," *Genome Research*, vol. 19, no. 2, pp. 266–275, 2009.

[9] L. Weis and D. Reinberg, "Transcription by RNA polymerase II: initiator-directed formation of transcription-competent complexes," *The FASEB Journal*, vol. 6, no. 14, pp. 3300–3309, 1992.

[10] S. T. Smale and J. T. Kadonaga, "The RNA polymerase II core promoter," *Annual Review of Biochemistry*, vol. 72, pp. 449–479, 2003.

[11] J. Zeng, S. Zhu, and H. Yan, "Towards accurate human promoter recognition: a review of currently used sequence features and classification methods," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 498–508, 2009.

[12] N. I. Gershenzon and I. P. Ioshikhes, "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis," *Bioinformatics*, vol. 21, no. 8, pp. 1295–1300, 2005.

[13] D. Takai and P. A. Jones, "Comprehensive analysis of CpG islands in human chromosomes 21 and 22," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 6, pp. 3740–3745, 2002.

[14] M. Scherf, A. Klingenhoff, and T. Werner, "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach," *Journal of Molecular Biology*, vol. 297, no. 3, pp. 599–606, 2000.

[15] S. Fujii, H. Kono, S. Takenaka, N. Go, and A. Sarai, "Sequence-dependent DNA deformability studied using molecular dynamics simulations," *Nucleic Acids Research*, vol. 35, no. 18, pp. 6063–6074, 2007.

[16] Y. Fukue, N. Sumida, J.-I. Tanase, and T. Ohyama, "A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance," *Nucleic Acids Research*, vol. 33, no. 12, pp. 3821–3827, 2005.

[17] J. Gu, T. He, Y. Pei et al., "Primary transcripts and expressions of mammal intergenic microRNAs detected by mapping ESTs to their flanking sequences," *Mammalian Genome*, vol. 17, no. 10, pp. 1033–1041, 2006.

[18] X. Zhou, J. Ruan, G. Wang, and W. Zhang, "Characterization and identification of microRNA core promoters in four model species," *PLoS Computational Biology*, vol. 3, no. 3, pp. 0412–0423, 2007.

[19] H. K. Saini, S. Griffiths-Jones, and A. J. Enright, "Genomic analysis of human microRNA transcripts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 45, pp. 17719–17724, 2007.

[20] S. Fujita and H. Iba, "Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates," *Bioinformatics*, vol. 24, no. 3, pp. 303–308, 2008.

[21] A. Marson, S. S. Levine, M. F. Cole et al., "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells," *Cell*, vol. 134, no. 3, pp. 521–533, 2008.

[22] G. Wang, Y. Wang, C. Shen et al., "RNA Polymerase II Binding Patterns Reveal Genomic Regions Involved in MicroRNA Gene Regulation," *PLoS ONE*, vol. 5, no. 11, Article ID e13798, 2010.

[23] D. L. Corcoran, K. V. Pandit, B. Gordon, A. Bhattacharjee, N. Kaminski, and P. V. Benos, "Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data," *PLoS ONE*, vol. 4, no. 4, Article ID e5279, 2009.

[24] C.-H. Chien, Y.-M. Sun, W.-C. Chang et al., "Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data," *Nucleic Acids Research*, vol. 39, no. 21, pp. 9345–9356, 2011.

[25] A. Marsico, M. R. Huska, J. Lasserre et al., "PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs," *Genome Biology*, vol. 14, no. 8, article R84, 2013.

[26] Z. Xie, E. Allen, N. Fahlgren, A. Calamar, S. A. Givan, and J. C. Carrington, "Expression of Arabidopsis MIRNA genes," *Plant Physiology*, vol. 138, no. 4, pp. 2145–2154, 2005.

[27] X. Zhao, H. Zhang, and L. Li, "Identification and analysis of the proximal promoters of microRNA genes in Arabidopsis," *Genomics*, vol. 101, no. 3, pp. 187–194, 2013.

[28] J. Ernst, P. Kheradpour, T. S. Mikkelsen et al., "Systematic analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, no. 7345, pp. 43–49, 2011.

[29] J. Ernst and M. Kellis, "Discovery and characterization of chromatin states for systematic annotation of the human genome," *Nature Biotechnology*, vol. 28, no. 8, pp. 817–825, 2010.

[30] C. Luo, D. J. Sidote, Y. Zhang, R. A. Kerstetter, T. P. Michael, and E. Lam, "Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production," *Plant Journal*, vol. 73, no. 1, pp. 77–90, 2013.

[31] A. Barski, R. Jothi, S. Cuddapah et al., "Chromatin poises miRNA- and protein-coding genes for expression," *Genome Research*, vol. 19, no. 10, pp. 1742–1751, 2009.

[32] A. Kozomara and S. Griffiths-Jones, "MiRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, pp. D68–D73, 2014.

[33] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[34] G. Robertson, M. Hirst, M. Bainbridge et al., "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nature Methods*, vol. 4, no. 8, pp. 651–657, 2007.

[35] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[36] J. Rozowsky, G. Euskirchen, R. K. Auerbach et al., "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls," *Nature Biotechnology*, vol. 27, no. 1, pp. 66–75, 2009.

[37] A. Barski, S. Cuddapah, K. Cui et al., "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, no. 4, pp. 823–837, 2007.

[38] B. Rhead, D. Karolchik, R. M. Kuhn et al., "The UCSC genome browser database: update 2010," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D613–D619, 2009.

[39] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[40] T. S. Mikkelsen, M. Ku, D. B. Jaffe et al., "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, no. 7153, pp. 553–560, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 238139, 8 pages
http://dx.doi.org/10.1155/2015/238139

Research Article

Constructing a Genome-Wide LD Map of Wild *A. gambiae* Using Next-Generation Sequencing

Xiaohong Wang,[1] Yaw A. Afrane,[2] Guiyun Yan,[3] and Jun Li[1]

[1]*Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK 73019, USA*
[2]*Centre for Global Health Research, Kenya Medical Research Institute, Kisumu 40100, Kenya*
[3]*Program in Public Health, University of California, Irvine, CA 92697, USA*

Correspondence should be addressed to Jun Li; junli@ou.edu

Received 3 December 2014; Accepted 24 February 2015

Academic Editor: Guohua Wang

Anopheles gambiae is the major malaria vector in Africa. Examining the molecular basis of *A. gambiae* traits requires knowledge of both genetic variation and genome-wide linkage disequilibrium (LD) map of wild *A. gambiae* populations from malaria-endemic areas. We sequenced the genomes of nine wild *A. gambiae* mosquitoes individually using next-generation sequencing technologies and detected 2,219,815 common single nucleotide polymorphisms (SNPs), 88% of which are novel. SNPs are not evenly distributed across *A. gambiae* chromosomes. The low SNP-frequency regions overlay heterochromatin and chromosome inversion domains, consistent with the lower recombinant rates at these regions. Nearly one million SNPs that were genotyped correctly in all individual mosquitoes with 99.6% confidence were extracted from these high-throughput sequencing data. Based on these SNP genotypes, we constructed a genome-wide LD map for wild *A. gambiae* from malaria-endemic areas in Kenya and made it available through a public Website. The average size of LD blocks is less than 40 bp, and several large LD blocks were also discovered clustered around the *para* gene, which is consistent with the effect of insecticide selective sweeps. The SNPs and the LD map will be valuable resources for scientific communities to dissect the *A. gambiae* genome.

1. Background

Malaria, a mosquito-transmitted disease caused by parasites of the genus *Plasmodium*, leads to as many as 300 million clinical cases per year [1]. Of these, approximately one million die from malaria, with 75% of the deaths occurring in African children. Human malaria parasites are transmitted by anopheline mosquitoes, of which *Anopheles gambiae* is the most prevalent vector in Africa.

Genetic variation in mosquito populations affects the mosquitoes' susceptibility to *P. falciparum* infection [2–4], insecticide resistance [5–8], and other traits of interest. Determining the molecular basis for these and other important mosquito traits requires knowledge of genome-wide genetic variation and high-resolution linkage maps in wild *A. gambiae* populations from malaria-endemic areas. The currently available *A. gambiae* SNPs in the NCBI database dbSNP mainly derive from laboratory mosquito colonies [9, 10]. Sampling a small set of genes [11] or SNPs [12] in field-collected samples indicated low linkage disequilibrium (LD) in *A. gambiae* populations. However, this result means neither that neighboring SNPs are not linked, nor that large LD blocks in the *A. gambiae* genome do not exist. Therefore, it is still critical to define the extent of genome-wide genetic variation and linkage information in *A. gambiae* populations from malaria-endemic areas. Recent advances in sequencing technologies and bioinformatics make it economically feasible for a single research laboratory to detect genome-wide SNPs and to construct an *A. gambiae* LD map using wild-derived mosquitoes. Depending on needs, the current available software such as Haploview [13] allows the users to easily generate an interactive haplotype map for a whole genome or a certain genomic region based on a set of genotypes or LD map.

In this study, we collected wild *A. gambiae* from Kenya, sequenced individual mosquitoes with Illumina sequencing

technologies, developed novel pipelines to detect SNPs, constructed an *A. gambiae* LD map, and established a computer server to present the data to the public. Notably, the consistence between our data and experimental findings supports the accuracy of this resource and demonstrates the advantages of the SNPs and LD map.

2. Results

2.1. Detecting SNPs in Wild A. gambiae Mosquitoes. To detect common SNPs (frequencies > 5%) in wild *A. gambiae* populations, individual genomic DNA from nine randomly selected wild *A. gambiae* mosquitoes recovered from highland areas around Kisumu was studied. DNA from each mosquito was sequenced individually after the samples were confirmed to be *A. gambiae*. Each sequence read was 100 bp long. The average sequencing coverage of the whole genome for each individual was greater than 36.1-fold. These reads were mapped onto the *A. gambiae* reference genome [9]. More than 1.6 million SNPs were detected in each individual mosquito according to the aligned short reads. Among the detected SNPs, 2,219,815 common SNPs were detected in more than one mosquito, and about 4,911,116 unique SNPs were detected in only one mosquito. The SNP-frequency is about one SNP per 33 bp, which is consistent with previous reports [11, 14]. To check the detection accuracy, a randomly selected set of SNPs was verified using a graphical user interface (http://omics.ou.edu/AgHapMap). The results indicated a low error rate of less than 0.1%. Notably, as many as 87.6% of the newly discovered SNPs are novel, compared to SNPs in dbSNP (NCBI, release 125) [15] that were mainly from mosquito colonies maintained in laboratories. This suggests that these SNPs are useful resources to study *A. gambiae* in the field. Among the novel common SNPs, 36,675 (1.9%) are nonsynonymous, 135,500 (7.0%) are synonymous, 371,417 (19.1%) are within intron regions, and the others (1,401,750, 71.3%) are at intergenic regions (Figure 1). The SNP type distribution for novel SNPs is similar to the known SNPs.

We identified that four large genomic regions, for example, 2R:57.6 MB-2L:4.0 MB, 2L:20.5–42.2 MB, X:17.6–24.4 MB, and 3R:52.0 MB-3L:0.4 MB, have much lower frequencies of SNPs, compared to other regions (Figure 2(a)). As expected, three of these regions (labeled with green lines in Figure 2(a)) are around centromeres on chromosome 2, X and 3, consistent with characteristics of heterochromatic regions [16]. Strikingly, a region on chromosome arm 2L from 20.5 MB to 42.2 MB (labeled with red line in Figure 2(a)) also exhibited low frequencies of SNPs and overlaid a chromosomal inversion called 2La [17]. We karyotyped the inversion of 2La in the nine mosquitoes using PCR [18], and the results show that four mosquitoes were $2L+^{a}/2L+^{a}$ and five were 2La/2La. The 2L chromosomal inversion region always had fewer SNPs than other genomic regions, regardless of mosquito karotypes. The consistence of the low recombinant rate at heterochromatic regions and chromosomal inversion to the SNP distribition partially validates the SNPs genome-widely.

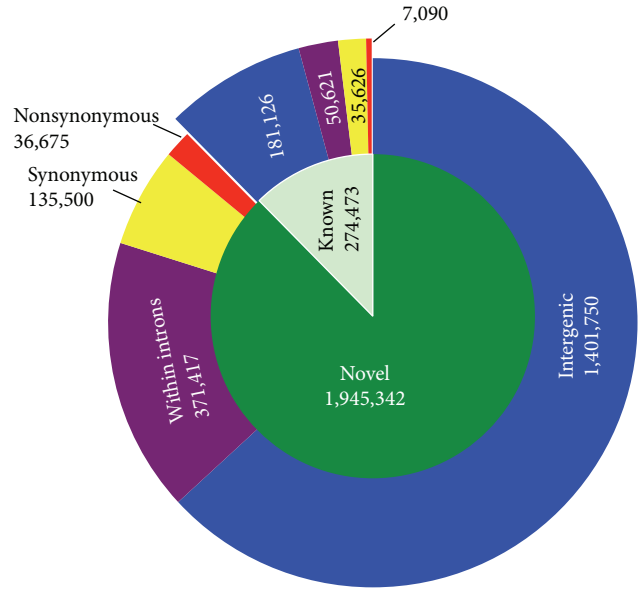


Figure 1: Types of common SNPs. About 87.6% of newly detected common SNPs from wild *A. gambiae* are novel. The types of common SNPs were determined based on their positions on the genome (intergenic: blue; within introns: purple; synonymous: yellow; nonsynonymous: red). SNPs within exons were further classified into synonymous and nonsynonymous. About 2% of SNPs changed protein sequences. The number of SNPs is shown in each category.

2.2. The LD Map in Wild A. gambiae Populations from Malaria-Endemic Areas in Kenya. We next extracted the SNP genotypes of individual mosquitoes based on the high-throughput short read sequences as described in the Methods section. Out of 2,219,815 common SNPs, 785,687 SNPs (one SNP per every 293 bp genome-wide) were reliably genotyped at 99.6% confidence in nine mosquitoes. The correlation coefficient among SNPs was calculated by using Haploview software [13]. As shown in Figure 3(a), the average coefficient of determination over distance between SNPs decreases rapidly. Moreover, when the distance between two SNPs is greater than 40 bp, the linkage relation between SNPs is nearly random (Figure 3(b)).

Although the average LD size in *A. gambiae* is very short (Figure 3), which is consistent with previous reports [11, 12], our genome-wide, high-throughput LD analysis also identified regions with very large LD blocks (Figures 2(b) and 2(c)). For instance, the locus at 2L: 1.8 MB–4.2 MB contains four large LD blocks, and the average LD size at locus 2L: 20.5 MB–22.6 MB is apparently greater than the LD size at locus 2L: 47.1 MB–48.3 MB (Figure 2(b)). To accurately quantitate the linkage relationship between neighboring SNPs efficiently, we calculated the number of SNPs per LD block. According to the plot of coefficient of determination versus distance (Figure 3(b)), two neighboring SNPs were considered to be linked if their correlation coefficient was greater than 0.25. Although most of LD blocks contained less than 3 SNPs, several very large LD blocks with more than 50 SNPs

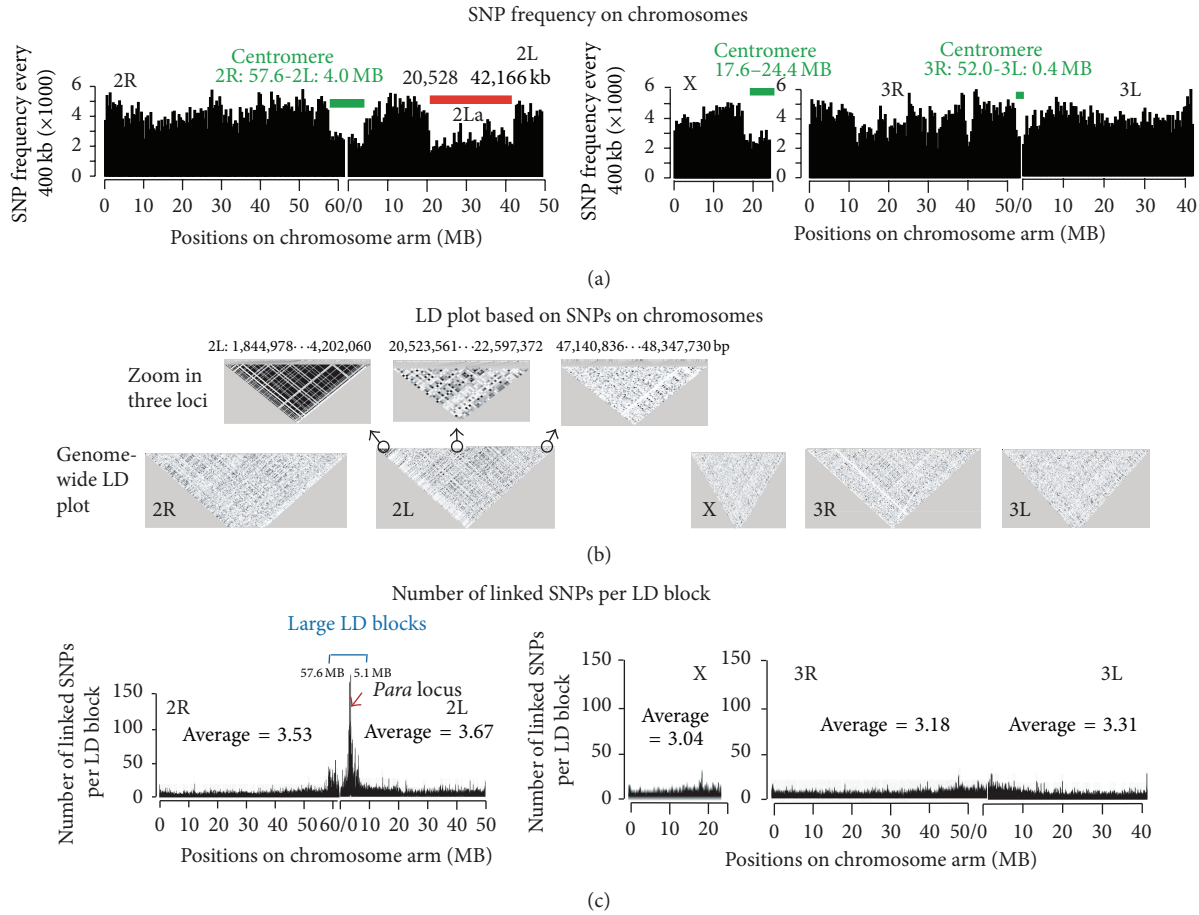


Figure 2: Genome-wide SNP-frequency, LD plots, and number of SNPs per LD block. (a) shows the SNP-frequency (per 400 kb) on *A. gambiae* chromosomes. The genomic regions (>1 MB) with less than 1 SNP per 150 bp were labeled with green and red lines. (b) shows LD plot of five chromosome arms and zoom in of three particular regions (high LD, chromosome inversion region, and other regions) to illustrate the genome-wide linkage map in detail. (c) The number of SNPs per LD block on chromosomes. The x-axes of (a) and (c) correspond to the same positions.

were clustered at a locus on chromosome 2 (2R:57.6 MB-2L:5.1 MB), indicated by a blue line in Figure 2(c). Detailed analysis of the genes within this genomic region clustering large LD blocks found that the *para* gene (AGAP004707) was at the center (2L, 2.4 MB) of the large LD clusters (Figure 2(c)). All nine sequenced mosquitoes from malaria-endemic areas at highland areas around the Kisumu district in western Kenya are homozygous for the insecticide resistant allele cytosine, which forms the code of "TCA" and encodes amino acid serine in the voltage-gated sodium channel [19]. The biological reason of this locus validates the new LD map and demonstrates the usability of the LD map.

We further verified the *A. gambiae* LD map experimentally. Two pairs of SNPs within neighboring genes were genotyped in 22 randomly selected female wild-derived *A. gambiae*: SNP at chromosomal arm 2L, 39,852,810 bp within gene AGAP006906 versus SNP at chromosomal arm 2L, 39,966,795 bp within gene AGAP006914, and SNP at chromosomal arm 2L at 41,165,983 bp within gene AGAP007031 versus SNP at chromosomal arm 2L at 41,246,582 bp within gene AGAP007032. The results (Table 1) showed that the coefficient of determination between pairs of SNPs was less than 0.05, experimentally validating the computational LD results from high-throughput sequencing.

Table 1: Correlation coefficient between nonsynonymous SNPs of two pairs of neighboring genes.

	Based on HT (9 individuals)	Based on clone (22 individuals)
AGAP006906 versus AGAP006914	0.156	0.009
AGAP006914 versus AGAP007031	0.156	0.02
AGAP007031 versus AGAP007032	0.044	<0.001

AGAP006906, SNP position (bp), 39852810; AGAP006914, SNP position, 39966795; AGAP007031, SNP position, 41165983; AGAP007032, SNP position, 41246582. HT: high-throughput sequencing data. Clone: PCR fragments from individual mosquitoes.

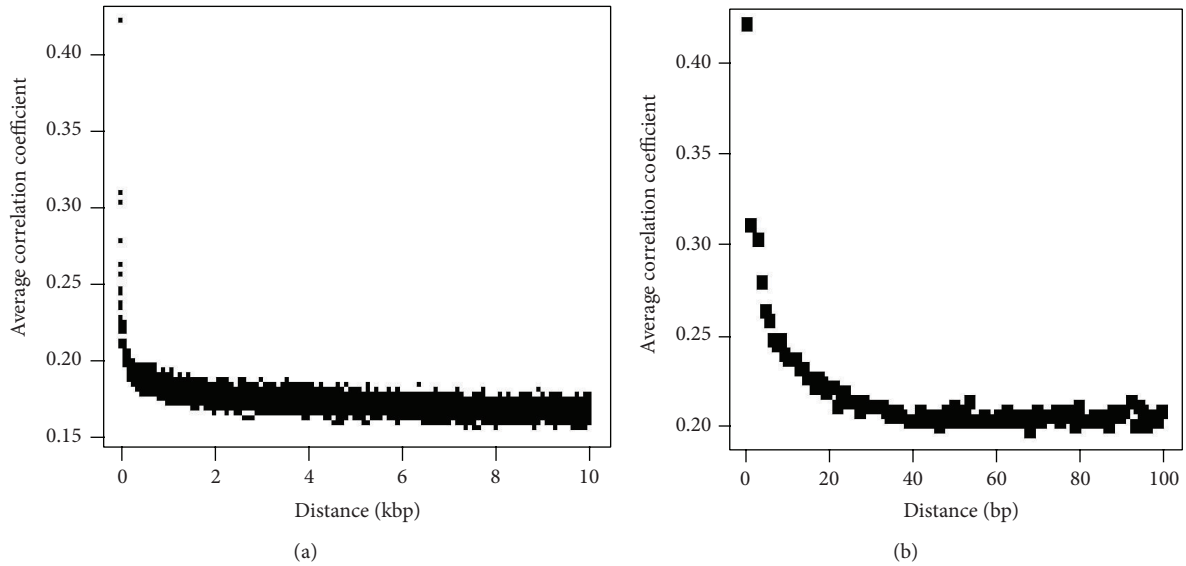


Figure 3: LD decays rapidly as the distance of SNPs increases. (a) displays the relationship between correlation coefficient and SNP distance from 0 to 10 kb. (b) shows the relationship between correlation coefficient and SNP distance from 0 to 100 bp, which clearly shows that average genome-wide LD size is less than 40 bp.

2.3. Public Web to Integrate Aligned Reads, SNPs, LD Map, and Genome Annotation. To make these valuable data available to the scientific community, we established a computer server and constructed databases and a Web interface to visualize the SNPs, LD map, short reads, and detailed alignments, along with internal and external genome annotations. The Website is accessible through http://omics.ou.edu/AgHapMap. After access, users can click on the tab "Select Tracks" to select the data that they are interested in and click on "Browser" to see actual data. To zoom in on a particular region, they can highlight the region and click on "zoom in." Figure 4 shows the screen shot of this server.

3. Discussion

Genetic variation and LD maps are two important resources that enable identification of genetic mutants associated with traits of interest in populations. However, it is impractical to detect genetic variation and build a genome-wide LD map for all species with traditional approaches, for example, by surveying a set of genetic markers in populations. To overcome these limitations, we extracted and sequenced genomic DNA from individual mosquitoes with high-throughput sequencing technologies. Next, we developed a pipeline to obtain more than two million SNPs. Importantly, the majority (88%) of our SNPs from wild-derived *A. gambiae* are novel, which will help the community to address the molecular mechanisms of trait determination, as well as potentially reconciling discrepancies when comparing results obtained from laboratory mosquitoes versus field isolates [20]. Furthermore, we developed a novel computational approach to genotype nearly one million SNPs in individual mosquitoes solely based on our high-throughput sequencing data. Traditional approaches for genotyping individuals requires a *priori* knowledge of genetic markers, and it is tedious to genotype each genetic marker in individuals using hybridization-based methods or PCR-based approaches. In this report, we combined SNP discovery with SNP genotyping using a new computational pipeline, making the process both more efficient and cost-effective. Using our high-throughput sequencing data and our computational pipeline, we have assembled the first genome-wide LD map of *A. gambiae,* using wild-derived mosquitoes from malaria-endemic areas in western Kenya. We also report here that the newly detected genetic variation and LD map have been made freely and easily accessible to the public through the Internet. Notably, our approach and pipeline are applicable to generate LD maps of other biologically, agriculturally, and medically important mosquito species.

As mentioned above, it is well-known that the genetic variation and LD maps are important for association studies to analyze wild mosquitoes [2, 3], and we have also demonstrated their utility in our previous publication [4]. The interaction among multiple genetic variation within multiple genes also contributes to a complex trait [21]. Here, we highlight additional and powerful applications of genetic variation and the LD map to investigate aspects of mosquito biology in nature.

SNP distribution identified four large genomic regions harboring unusually low frequencies of SNPs, three of which localized to centromeres. Although this observation was expected, given that the genomic regions around centromeres have lower recombination rates than other loci and DNA recombinant rates and SNP density are positively correlated [22], we also identified a fourth large locus with lower SNP-frequency on 2L at 20.5–42.2 MB. Notably, the fourth locus was also associated with lower recombination rates because this region contains a chromosomal inversion [17],

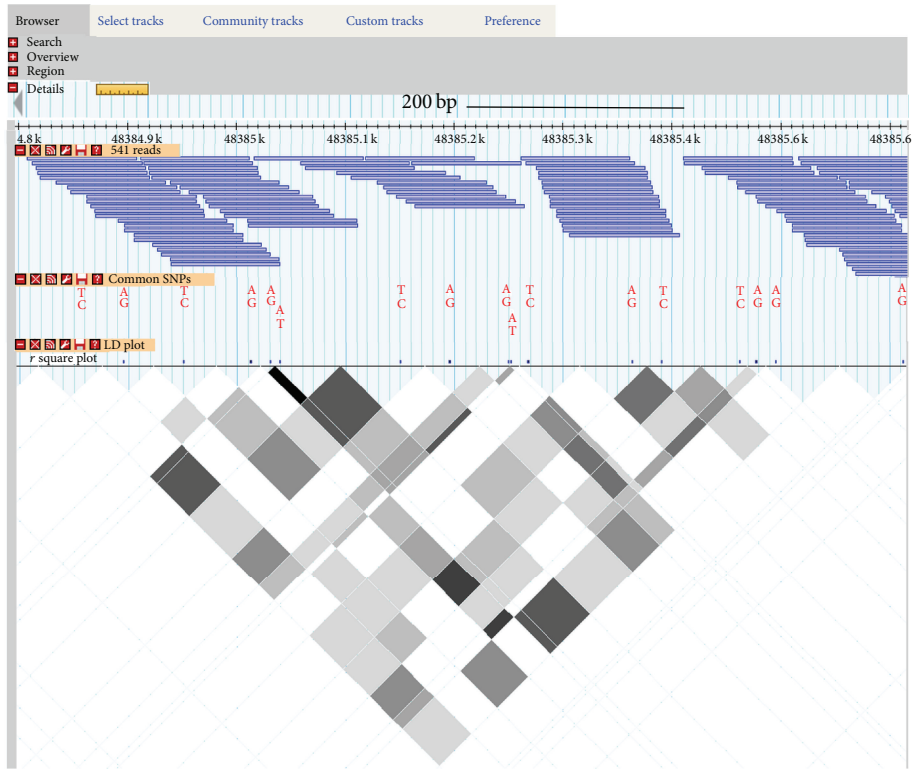


Figure 4: Screenshot of the Web interface to display SNPs, reads, LD Map, and external genome annotation. Users can obtain reads, SNPs, and LD at http://omics.ou.edu/AgHapMap. The sequences and alignments of reads can be viewed in detail by highlighting and zooming in. Data tracks from internal and external databases, which are not shown in this screenshot, can be integrated by selecting through the Tab of "Select Tracks."

and chromosomal inversions are known to inhibit DNA recombination [23]. We karyotyped the inversion forms of 2La in individual *A. gambiae* to investigate whether our sample contained a single karyotype form that causes the lower SNP-frequency at this region. Our results show that the lower SNP-frequency did not associate with any particular karyotype form of chromosomal inversion of 2La. Apparently, the consistence between SNP distribution and genetic data validates the detected SNPs genome-widely and demonstrates their usability.

Regarding our *A. gambiae* LD map, a previous survey of a limited set of genes ($n = 4$) [11] or SNPs ($n = 1{,}536$) [12] suggested low LD (<200 bp) in *A. gambiae* populations. Our data are consistent with those reports. However, here we extend these observations and show that the average LD size in *A. gambiae* populations in western Kenya is less than 40 bp. It is worth noting that our new results additionally provide a genome-wide LD map of nearly one million genetic markers. Furthermore, the LD map reveals a genomic locus on chromosome 2 (2R: 57.6 MB-2L: 5.1 MB) that is clustered with larger LD blocks. Analyses of the genes within this region identified that the *para* gene is at the center of this locus. The *para* gene encodes a voltage-gated sodium channel (VGSC) that is the target molecule of common insecticides such as pyrethroids [24]. The well-known *kdr* mutations, which change codon 1014 from leucine to serine or phenylalanine within the *para* gene coding region, confer insecticide resistance [25–27]. Indeed, all mosquitoes that we sequenced ($n = 9$) harbored the resistance allele (1014S) instead of the wild type allele (1014L). It is well known that the use of insecticides remains the traditional approach to combating the spread of malaria [28]. The molecular target of common insecticides such as pyrethroids and dichlorodiphenyltrichloroethane (DTT) is VGSC [24, 29]. DDT and pyrethroids were used globally, including Kenya [30], and caused an insecticide-driven selective sweeping in western Kenya. These data are consistent with insecticide resistance bioassays in the field [19, 31] where our mosquitoes were sampled. Rapid rise of *kdr* mutation frequency and even fixation over the past decade when pyrethroid insecticides have been used extensively in Africa suggest the importance of this mechanism in the process of pyrethroid resistance. On the other hand, given the fixation of *kdr* mutations in many *A. gambiae* populations, metabolic detoxification is becoming an increasingly important resistance mechanism. Clearly, vector insecticide resistance is an outstanding issue in the control and prevention of vector-borne diseases as supported by our data and other reports [32, 33]. Collectively, the larger LD around the *para* gene validates our LD map and demonstrates an application of using our LD map to detect genomic regions under selection pressure.

In conclusion, we collected and sequenced wild *A. gambiae* mosquitoes from malaria-endemic areas in Kenya using next-generation sequencing technology and developed

a pipeline to analyze SNPs and genotypes. More than 2 million common SNPs were identified in wild *A. gambiae* populations, and 785,687 SNPs were genotyped in nine mosquitoes. Using these data, we constructed the first genome-wide *A. gambiae* LD map, which will serve as a powerful and useful resource to dissect the mosquito genome. The consistence between our data and previous findings supports the accuracy of this resource.

4. Methods

4.1. Sampling Wild A. gambiae. Collecting and rearing mosquitoes were performed as described previously [4]. In brief, *A. gambiae* larvae were collected from natural habitats (>10 meters distance between any two habitats) in highland areas around the Kisumu district of Kenya where malaria is hyperendemic. More than half of mosquito larvae were successfully reared to adults in an insectary at the Kenya Medical Research Institute. The resulting 3–5-day post-emergence female mosquitoes were used for experiments. It is worth noting that only the female wild-derived mosquitoes that fed on human blood through membrane feeding were further analyzed in this study. Genomic DNA was extracted from 7-day post-blood-fed mosquitoes using DNAzol (Life Technologies, Grand Island, NY, USA). The individual mosquito species was confirmed by the rDNA-PCR method [34].

4.2. Sequencing Individual A. gambiae Genomes and Detecting SNPs. Genomic DNA from individual mosquitoes was sheared to construct a DNA library with fragment lengths of about 300 bp, and both sides of each DNA fragment were sequenced in lengths of 100 bp. These reads were mapped to the *A. gambiae* reference genome (assembly version AgamP3) using the short oligonucleotide analysis package (SOAP) [35]. SOAP used the seed-and-hash algorithm to align high-throughput sequences onto the reference genome accurately and efficiently. To focus on the SNP detection, we turned off the option for gaps, for example, "soap –a leftReads –b rightRead –D referenceGenome.index –o AlignedFile.txt –m 50 –x 550 –g 0." The alignments for each chromosome were then extracted using the linux command "grep," for example, "grep X AlignedFile.txt > X.align," followed by sorting the output based on alignment position on chromosome, for example, "sort –k9 –n X.align > X.align.sort." Finally, the "soapsnp" program in the SOAP package was used to detect nucleotide variation at each position, for example, "soapsnp –i X.align.sort –d reference_genome_seq.fasta –o SNPonX." At each genome position, we extracted the SNPs that had at least one uniquely mapped read for best base and at least one uniquely mapped read for second best base (phrep score > 30) [36]. The nucleotides that are different from the reference sequence were also extracted into the SNP set. To obtain common SNPs, we removed the SNPs that (1) were detected in only one mosquito and (2) were identical in all nine mosquitoes (they were detected because they were different from the reference genome). Finally, we checked the error rate by using our web interface. We randomly selected SNPs three times with 100 SNPs each time. Then we manually examined the aligned sequence reads to the reference genome sequences to count the true positives.

Table 2: Primers to clone two pairs of neighboring genes.

AGAP006906	Forward	5′-CGGAGGCACACACCATCA-3′
	Reverse	5′-GCGAAACTCCAGATACAGCA-3′
AGAP006914	Forward	5′-CAACTGCTGGCCAAAGGAC-3′
	Reverse	5′-GTCCTTTGGCCAGCAGTTG-3′
AGAP007031	Forward	5′-GGCTCGAAGTCCGATTACA-3′
	Reverse	5′-GTCGGCACAGTCGTGGTA-3′
AGAP007032	Forward	5′-ATAACCATGCGGAGAGTGTG-3′
	Reverse	5′-CCGTTCGATTTCCTCCTG-3′

4.3. Genotyping the SNPs in Nine Individual Mosquitoes Based on High-Throughput Sequencing Data. For each detected SNP, we checked all reads in each individual mosquito regardless of its sequencing quality score (Phred score) [36]. If two alleles for an SNP were detected in reads from one *A. gambiae* individual, a heterozygous genotype was assigned to that individual for that genome position. However, for each homozygous SNP genotype, the number of reads hitting that position was counted using the information obtained through the program "soapsnp" from the SOAP package [35]. If a homogenous SNP genotype was supported by at least eight reads from that mosquito, it was kept for LD map analysis, because the *P* value of missing a heterozygous allele is less than 0.004 based on the binomial distribution. The genotypes of these SNPs were applied to calculate the LD among SNPs by using the software Haploview [13]. Two neighboring SNPs with a correlation coefficient greater than a threshold (e.g., 0.25 in Figure 3(b)) were treated as linked SNPs in one LD block.

4.4. Genotyping a Particular Set of SNPs in Individual Mosquitoes Using PCR Followed by the Sanger Sequencing. We cloned two pairs of neighboring genes: AGAP006906 versus AGAP006914 and AGAP007031 versus AGAP007032 to verify the LD map. The sequences around nonsynonymous SNPs were cloned from 22 randomly selected female wild-derived *A. gambiae* by PCR with primers shown in Table 2. The PCR products were purified using QIAGEN PCR purification kits. The purified DNA fragments were sequenced with one PCR primer using Sanger approach. The sequencing trace files were displayed using software 4Peaks (http://nucleobytes.com/), and the SNPs were read manually.

4.5. Visualizing SNPs, Short Reads, Genome Annotation, and the HapMap through an Integrated Web Interface. The individual short reads, the sequence alignment to the reference genome, and the SNPs were compiled into databases as instructed by Gbrowse and displayed as tracks [37]. In brief, "samtools" (Sequence Alignment/Map tools) software downloaded from http://samtools.sourceforge.net/ was used to transform a data file from one format to another. The "samtool" was also used to import data into Gbrowse required databases. To generate these databases, the reference

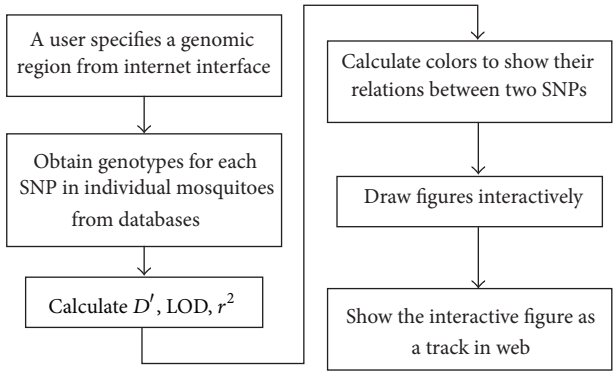


Figure 5: Protocol to construct a Web server to display *A. gambiae* LD Map.

genome sequence was indexed (e.g., "samtools faidx ReferenceSequence.fa"), and then the data were imported into databases (e.g., "samtool import ReferenceSequence.fa.fai"). The alignment files that contain short reads aligned to the reference sequence were sorted and indexed (e.g., "samtool sort aligmentFile.bam aligmentFile.sorted.bam" and "samtool index aligmentFile.sorted.bam") sequentially. Finally, the sorted and indexed alignment files were imported into databases (e.g., "samtool import aligmentFile.sorted.bam.bai"). SNP data were stored in a mysql database with a single table that was created with this command: "CREATE TABLE Agam_common_snps_position (snp varchar(10) NOT NULL, alleles varchar(4) NOT NULL, chr ENUM ('2L', '2R','3L','3R','X'), pos int(10) unsigned NOT NULL default '0', Ag541 char(2), Ag544 char(2), Ag545 char(2), Ag551 char(2), Ag553 char(2), Ag564 char(2), Ag565 char(2), Ag566 char(2), Ag567 char(2), PRIMARY KEY (chr,pos), KEY chr (chr), KEY pos (pos));". Each data set was displayed as a track (also known as a "plug-in") through Gbrowse. For instance, the genome annotation, including gene structure predictions in our internal databases (ReAno) [10] and external databases of https://www.vectorbase.org/ [38], was integrated as two tracks on the Web. The correlation-coefficient values among SNPs were constructed and integrated into the Web interface. To display an interactive graphic of the linkage map through the Internet, we constructed the server using a protocol as shown in Figure 5. In brief, the server obtains the interactive coordinators of SNPs and SNP genotypes from the databases and calculates the LD (D'), logarithm of odds (LOD), and coefficient of determination (r^2) for each pair of SNPs interactively. Based on a user's HapMap configuration (or default), the server calculates the colors and displays it as a plug-in track on the Web interface. Using downloaded Haploview software [13], users can easily generate an interactive haplotype map at areas of interest (such as large LD blocks) by highlighting the LD blocks with the mouse under the "LD plot" tab and then clicking on the tab "Haplotypes." The phased haplotypes of the highlighted blocks will then be displayed. The constructed HapMap may need further experimental validation.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contributions

Xiaohong Wang analyzed the sequence and constructed the databases, Web server, and interface. Yaw A. Afrane and Guiyun Yan participated in sampling wild *A. gambiae*. Jun Li designed research, analyzed the data, and wrote the paper. All authors participated in editing.

Acknowledgments

This work was partially supported by NIH/NIAID (1R56AI081829) and Oklahoma Center for Advancement of Science and Technology (HR13-055) to Jun Li. The authors also thank Dr. Noah Butler from University of Oklahoma Health Sciences Center for reading the whole paper and providing insights and comments.

References

[1] R. E. Cibulskis, M. Aregawi, R. Williams, M. Otten, and C. Dye, "Worldwide incidence of malaria in 2009: estimates, time trends, and a critique of methods," *PLoS Medicine*, vol. 8, no. 12, Article ID e1001142, 2011.

[2] D. M. Menge, D. Zhong, T. Guda et al., "Quantitative trait loci controlling refractoriness to *Plasmodium falciparum* in natural *Anopheles gambiae* mosquitoes from a malaria-endemic region in western Kenya," *Genetics*, vol. 173, no. 1, pp. 235–241, 2006.

[3] M. M. Riehle, K. Markianos, O. Niaré et al., "Natural malaria infection in *Anopheles gambiae* is regulated by a single genomic control region," *Science*, vol. 312, no. 5773, pp. 577–579, 2006.

[4] J. Li, X. Wang, G. Zhang, J. I. Githure, G. Yan, and A. A. James, "Genome-block expression-assisted association studies discover malaria resistance genes in *Anopheles gambiae*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 51, pp. 20675–20680, 2013.

[5] M. Bonizzoni, Y. Afrane, W. A. Dunn et al., "Comparative transcriptome analyses of deltamethrin-resistant and -susceptible *Anopheles gambiae* mosquitoes from Kenya by RNA-Seq," *PLoS ONE*, vol. 7, no. 9, Article ID e44607, 2012.

[6] J. M. Riveron, H. Irving, M. Ndula et al., "Directionally selected cytochrome P450 alleles are driving the spread of pyrethroid resistance in the major malaria vector Anopheles funestus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 1, pp. 252–257, 2013.

[7] C. V. Edi, L. Djogbénou, A. M. Jenkins et al., "CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito *Anopheles gambiae*," *PLoS Genetics*, vol. 10, no. 3, Article ID e1004236, 2014.

[8] B. Kabula, W. Kisinza, P. Tungu et al., "Co-occurrence and distribution of East (L1014S) and West (L1014F) African knockdown resistance in *Anopheles gambiae* sensu lato population of Tanzania," *Tropical Medicine and International Health*, vol. 19, no. 3, pp. 331–341, 2014.

[9] R. A. Holt, G. Mani Subramanian, A. Halpern et al., "The genome sequence of the malaria mosquito *Anopheles gambiae*," *Science*, vol. 298, no. 5591, pp. 129–149, 2002.

[10] J. Li, J. M. C. Ribeiro, and G. Yan, "Allelic gene structure variations in *Anopheles gambiae*," *PLoS ONE*, vol. 5, no. 5, Article ID e10699, 2010.

[11] C. Harris, F. Rousset, I. Morlais, D. Fontenille, and A. Cohuet, "Low linkage disequilibrium in wild *Anopheles gambiae s.l.* populations," *BMC Genetics*, vol. 11, article no. 81, 2010.

[12] D. Weetman, C. S. Wilding, K. Steen, J. C. Morgan, F. Simard, and M. J. Donnelly, "Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: major variants identified in a low-linkage disequilbrium genome," *PLoS ONE*, vol. 5, no. 10, Article ID e13140, 2010.

[13] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.

[14] C. S. Wilding, D. Weetman, K. Steen, and M. J. Donnelly, "High, clustered, nucleotide diversity in the genome of Anopheles gambiae revealed through pooled-template sequencing: implications for high-throughput genotyping protocols," *BMC Genomics*, vol. 10, article 320, 2009.

[15] S. F. Saccone, J. Quan, G. Mehta et al., "New tools and methods for direct programmatic access to the dbSNP relational database," *Nucleic Acids Research*, vol. 39, no. 1, pp. D901–D907, 2011.

[16] C. Cheng, B. J. White, C. Kamdem et al., "Ecological genomics of anopheles gambiae along a latitudinal cline: a population-resequencing approach," *Genetics*, vol. 190, no. 4, pp. 1417–1432, 2012.

[17] I. V. Sharakhov, B. J. White, M. V. Sharakhova et al., "Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the *Anopheles gambiae* complex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 16, pp. 6258–6262, 2006.

[18] B. J. White, F. Santolamazza, L. Kamau et al., "Molecular karyotyping of the 2La inversion in *Anopheles gambiae*," *The American Journal of Tropical Medicine and Hygiene*, vol. 76, no. 2, pp. 334–339, 2007.

[19] D. K. Mathias, E. Ochomo, F. Atieli et al., "Spatial and temporal variation in the kdr allele L1014S in *Anopheles gambiae* s.s. and phenotypic variability in susceptibility to insecticides in Western Kenya," *Malaria Journal*, vol. 10, article 10, 2011.

[20] K. Michel, C. Suwanchaichinda, I. Morlais et al., "Increased melanizing activity in *Anopheles gambiae* does not affect development of *Plasmodium falciparum*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 45, pp. 16858–16863, 2006.

[21] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou, "An overview of SNP interactions in genome-wide association studies," *Briefings in Functional Genomics*, 2014.

[22] M. W. Nachman, "Single nucleotide polymorphisms and recombination rate in humans," *Trends in Genetics*, vol. 17, no. 9, pp. 481–485, 2001.

[23] M. Kirkpatrick, "How and why chromosome inversions evolve," *PLoS Biology*, vol. 8, no. 9, Article ID e1000501, 2010.

[24] H. Kawada, S. Z. M. Oo, S. Thaung et al., "Co-occurrence of point mutations in the voltage-gated sodium channel of pyrethroid-resistant aedes aegypti populations in Myanmar," *PLoS Neglected Tropical Diseases*, vol. 8, no. 7, Article ID e3032, 2014.

[25] K. Dong, "A single amino acid change in the para sodium channel protein is associated with knockdown-resistance (kdr) to pyrethroid insecticides in German cockroach," *Insect Biochemistry and Molecular Biology*, vol. 27, no. 2, pp. 93–100, 1997.

[26] D. Martinez-Torres, F. Chandre, M. S. Williamson et al., "Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector *Anopheles gambiae* s.s.," *Insect Molecular Biology*, vol. 7, no. 2, pp. 179–184, 1998.

[27] J. Pinto, A. Lynd, J. L. Vicente et al., "Multiple origins of knockdown resistance mutations in the afrotropical mosquito vector *Anopheles gambiae*," *PLoS ONE*, vol. 2, no. 11, Article ID e1243, 2007.

[28] J. A. Nájera, "Malaria control: achievements, problems and strategies," *Parassitologia*, vol. 43, no. 1-2, pp. 1–89, 2001.

[29] J. Hemingway, N. J. Hawkes, L. McCarroll, and H. Ranson, "The molecular basis of insecticide resistance in mosquitoes," *Insect Biochemistry and Molecular Biology*, vol. 34, no. 7, pp. 653–665, 2004.

[30] T. G. E. Davies, L. M. Field, P. N. R. Usherwood, and M. S. Williamson, "A comparative study of voltage-gated sodium channels in the Insecta: implications for pyrethroid resistance in *Anopheline* and other Neopteran species," *Insect Molecular Biology*, vol. 16, no. 3, pp. 361–375, 2007.

[31] L. Kamau, D. Agai, D. Matoke, L. Wachira, G. Gikandi, and J. M. Vulule, "Status of insecticide susceptibility in *Anopheles gambiae* sensu lato and *Anopheles funestus* mosquitoes from Western Kenya," *Journal of Insect Science*, vol. 8, article 11, 2008.

[32] J. Hemingway, L. Field, and J. Vontas, "An overview of insecticide resistance," *Science*, vol. 298, no. 5591, pp. 96–97, 2002.

[33] WHO, "Malaria Elimination: a field manual for lowland and moderate endemic countries," World Health Organization Report, WHO, Geneva, Switzerland, 2007.

[34] J. A. Scott, W. G. Brogdon, and F. H. Collins, "Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction," *The American Journal of Tropical Medicine and Hygiene*, vol. 49, no. 4, pp. 520–529, 1993.

[35] R. Li, C. Yu, Y. Li et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[36] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research*, vol. 8, no. 3, pp. 186–194, 1998.

[37] L. D. Stein, "Using GBrowse 2.0 to visualize and share next-generation sequence data," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 162–171, 2013.

[38] D. Lawson, P. Arensburger, P. Atkinson et al., "VectorBase: a data resource for invertebrate vector genomics," *Nucleic Acids Research*, vol. 37, no. 1, pp. D583–D587, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 831352, 9 pages
http://dx.doi.org/10.1155/2015/831352

Review Article

Survey of Programs Used to Detect Alternative Splicing Isoforms from Deep Sequencing Data *In Silico*

Feng Min, Sumei Wang, and Li Zhang

Department of Infectious Diseases, The Affiliated Chenggong Hospital of Xiamen University, The 174th Hospital of the Chinese People's Liberation Army, Xiamen, Fujian 361000, China

Correspondence should be addressed to Sumei Wang; wangsumei174@126.com

Received 26 November 2014; Revised 17 February 2015; Accepted 2 March 2015

Academic Editor: Yunlong Liu

Next-generation sequencing techniques have been rapidly emerging. However, the massive sequencing reads hide a great deal of unknown important information. Advances have enabled researchers to discover alternative splicing (AS) sites and isoforms using computational approaches instead of molecular experiments. Given the importance of AS for gene expression and protein diversity in eukaryotes, detecting alternative splicing and isoforms represents a hot topic in systems biology and epigenetics research. The computational methods applied to AS prediction have improved since the emergence of next-generation sequencing. In this study, we introduce state-of-the-art research on AS and then compare the research methods and software tools available for AS based on next-generation sequencing reads. Finally, we discuss the prospects of computational methods related to AS.

1. Introduction

Alternative splicing (AS) refers to the production of pre-mRNA via gene transcription to generate a number of mature mRNAs based on different splice modes, thereby increasing protein diversity. Since alternative splicing was discovered, studies have identified a large number of AS events in the human gene transcription process [1]. Based on high-throughput deep sequencing data, AS occurs in approximately 95% of the human genome [2]. AS is an important regulatory mechanism involved in the regulation of eukaryotic gene expression and proteome diversity [3]. The process is closely linked with many diseases, including cancer and diseases of the nervous system [4–6]. Thus, scholars in medicine, genetics, bioinformatics, and other fields have directed considerable research interest towards AS with the aim of identifying additional splicing events that could facilitate a deeper understanding of the AS regulatory mechanism.

Splice site recognition represents a key step in selective splicing research. Splice sites are used to predict the positions of exon/intron structures and splice site features, and splice site recognition is the traditional strategy used to predict alternative splice sites. Many algorithms, software, and databases for sequence alignment have emerged due to the application of first-generation sequencing. The research resources designed specifically for AS have gradually become richer, including a common ASD AS database [7]. However, the cost of first-generation sequencing is high; considerable efforts have been directed towards the goal of creating thousand- and hundred-dollar genome sequencing technology in the postgenomic era. Thus, the high throughput and low cost of next-generation sequencing technologies have provided a new stage for scientific research [8, 9].

AS was discovered in 1977 [10]. Subsequently, researchers realized the importance of AS due to its ability to regulate gene expression and facilitate protein diversity [11, 12]. The advantages of next-generation sequencing technology have opened a new stage of sequencing, and the study of the massive amounts of data generated by RNA-seq technology has become an important research direction.

RNA-seq (high-throughput RNA sequencing) represents a new method for the analysis of gene expression and transcriptomes. Many software tools and databases have appeared with the capacity to generate short sequence alignments and predictions on the basis of the alternative splice sites identified using RNA-seq.

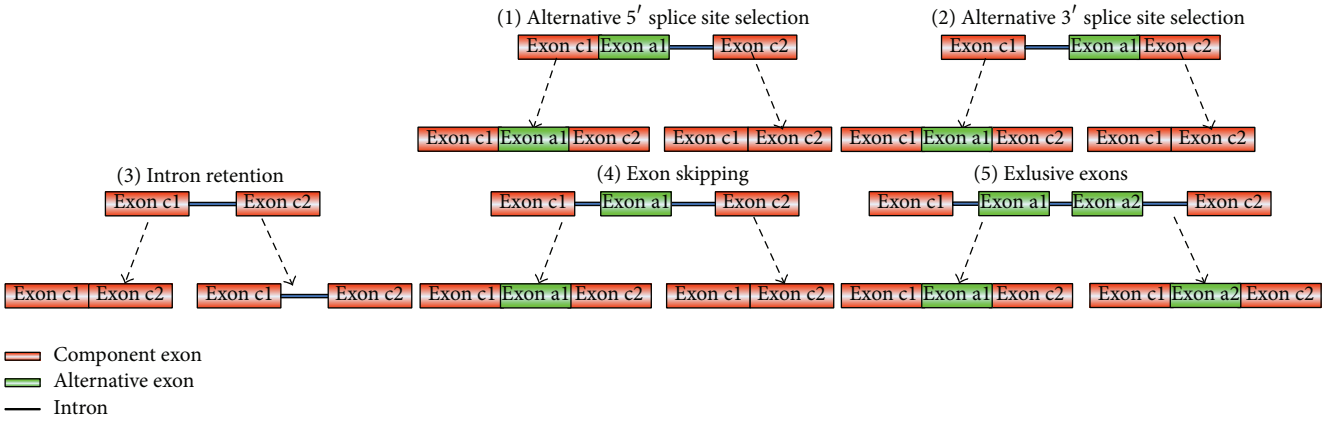


Figure 1: Five types of alternative splicing.

In this study, we outlined the methods, software tools, and databases available for AS research under two-generation sequencing technologies. The effect of these factors on AS research was analyzed. Using RNA-seq data produced by the Illumina/Solexa sequencing platform as an example, we compared three common splice site prediction programs (HMMSplicer [11], SOAPsplice, and TopHat [8]) under conditions of different depths and sequence read lengths. The performance of each type of software was evaluated under different conditions by comparing the number of accurately predicted sites, the accuracy rate, and the error rate. Finally, we discussed the problems and challenges associated with using deep sequencing data to study AS.

2. Discovering Alternative Splicing Sites from Long DNA Sequences

In addition to experimental methods, researchers predict potential AS events through the comparison between EST expression sequence tags and gene sequences. A large number of analyses and studies have validated the significance of the 3′ terminal splice acceptor site and 5′ terminal splice donor site in splicing events. Figure 1 summarizes the five AS forms.

The study by Fairbrother et al. [13] on exons in the human genome revealed that the splicing enhancers ESE and ESS serve an important regulatory function in selective splicing. Black [14] demonstrated that the splicing enhancer ISE and silencer ISS are also important for the selection of splicing sites and recognition of exons and introns. Thus, the AS process in eukaryotic genes is determined not only by a splicing factor but also by a complex regulatory process.

The means of selective splicing mainly include the following.

(1) Comparison analysis based on ESTs, mRNA, and gene fragments: EST comparative analysis was one of the earliest AS research methods. This method can identify certain AS events. However, EST has its own limitations, such as incomplete data, influence from genetic pollution, sensitive 3′ terminal, and high cost [15, 16]. Common comparison software programs include BLAT [17], Clustal [18], SIM4 [19], Ecgene [20], ASPIC [21], Spidey [22], GeneSeqer [23], and GMAP [24].

(2) Using gene chip high-throughput technology: gene chip technology has facilitated the research upsurge in the whole gene transcriptome. A large number of AS events have been identified using this technology. Johnson et al. [1, 25] discovered many exon-skipping events by analyzing microarray data. However, the disadvantage of this method is that probe density is limited, and designing a probe based on the known sequence and data analysis is difficult.

(3) Using machine learning methods for theoretical prediction: machine learning techniques have been widely used in various tasks in the field of bioinformatics, such as protein remote homology detection [26–29], microRNA identification [30, 31], protein binding site prediction [32], domain boundary identification [33, 34], DNA-binding protein prediction [35–37], protein structure prediction [38], enzyme classification [39, 40], gene regulation network construction [41], heat shock protein classification [42, 43], replication origin prediction [44, 45], nucleosome positioning sequence identification [46–48], CpG island methylation status prediction [49], translation initiation site prediction [50], promoter prediction [51], and microarray clustering [52, 53]. These machine learning based methods have achieved promising predictive performances. Therefore, some researchers have also applied common machine learning methods for theoretical predictions, such as support vector machine (SVM) [54, 55], weight matrices, the hidden Markov model, the quadratic discriminant function [56], and the neural network model [57]. The programs used for predicting splice sites based on these algorithms include HMMgene [58], NetGene2 [59, 60], geneID [61], GeneSplicer [62], and SpliceMachine [63].

3. Discovering Alternative Splicing Sites from Short Reads

The next-generation high-throughput sequencing technology developed rapidly after its emergence, thus enabling sequencing technology to move a step closer towards the thousand-dollar genome project. RNA-seq represents a new approach for gene expression and transcriptome studies. Currently, traditional AS research methods coexist with the development of the next-generation research methods. An increasing number of studies have been devoted to the development of new algorithms. In summary, next-generation high-throughput sequencing technology can provide a broad platform for AS due to its high efficiency and inexpensiveness.

However, RNA-seq also has shortcomings. The main challenge stems from read length. The read length of first-generation sequencing (i.e., Sanger sequencing) reaches approximately 1000 bp. The initial read length of RNA-seq was only approximately 25 bp. The read length is still relatively short, despite reaching 100 bp using Illumina/Solexa double-end sequencing [64].

3.1. Data Preprocessing. The first step in predicting an alternative splice site is to position the read on the reference transcriptome using RNA-seq data. However, the general analysis tools often position the reads on the reference genome because the transcriptome itself is not complete [8]. Short RNA-seq read lengths and incomplete transcriptomes cause the accuracy of this step to directly influence the accuracy of the prediction.

Some data found in read mapping can cross two exon-exon junctions [65]. This "read in junction" cannot be directly positioned on the genome sequence. This finding represents the key to studying alternative splice sites and identifying the critical region for exploring undetected splice events. Therefore, the processing strategy used to splice the read in junction is the key to predicting splice sites [66]. One approach for the treatment of read in junction is to position the reads onto the reference genome according to the currently known annotation of the exons. ERANGE [67] uses this method. Obviously, identifying new splice events is difficult using this approach. Another approach is to completely position the reads on the reference genome so that they can be divided into several different clusters. Reads with overlapping areas are classified into the same cluster. An exon region is delimited in each cluster [65]. Finally, the reads in junctions are positioned on the possible junctions. New splice events can be identified because the reads are based on known exon annotations. The splice site prediction software TopHat [8] uses this strategy.

Numerous software programs are specifically designed for the read mapping of RNA-seq data. These programs adopt the following algorithms: (1) the Smith-Waterman algorithm, such as BFAST [68] and SHRiMP [69]; (2) the two-way Burrows-Wheeler transform (BWT) algorithm, such as SOAPAligner [70]; (3) the BWT algorithm, such as Bowtie [71] and BWA [72]; and (4) the spaced-seed vacancy seed algorithm, such as MAQ [73]. Data compatibility should also be considered along with the choice of software. The formats of RNA-seq data generated by various sequencing platforms are different [74]. Thus, software versatility is affected by the styles and variety of formats it supports. Bowtie and BWA are relatively efficient, whereas SOAPAligner, BFAST, and MAQ have good tolerance for mismatches.

In addition to read mapping, we identified special software devoted to read assembly (i.e., de novo assembly). Few methods to study AS based on read assembly exist. However, read assembly has special roles in other biological information sciences. The typical read assembly software includes SHARCGS [75], SSAKE [76], and ALLPATHS [77]. The former two are assembled only for single sequence data, while the latter can be assembled for a pair of sequences from double-end sequencing. MAQ also has the ability to perform read assembly. Finally, sequence read archive (SRA) files are specialized for the storage of databases related to RNA-seq data for NCBI for inclusion into an AS database.

3.2. Alternative Splicing Prediction. The common AS site prediction software includes ERANGE, QPALMA [78], TopHat, MapSplice [79], SpliceMap, SOAPsplice, SplitSeek [80], and HMMSplicer. Current studies using RNA-seq to identify AS sites focus on locating splice sites, discovering new splice sites located as distantly as possible, and conducting next-step AS studies. Therefore, the accuracy and efficiency of predictions are key factors for the prediction software. Moreover, accuracy should be improved in order to predict more splice sites, while the error probability should be reduced; these factors differ for selected algorithms.

ERANGE was the earliest available method. It was the first program to use the read mapping method. In this method, the read is positioned on the reference genome based on known exon annotations. Thus, this method cannot be used to identify a new splice site. QPALMA adopts the machine learning strategy and trains support vector machines for site identification using known splice sites. Vmatch has been adopted for positioning. However, because the efficiency of Vmatch is not high enough compared with Bowtie, Vmatch is not used for comparing reads. TopHat first positions the sequence on the reference genome using Bowtie. MAQ successfully positions the sequence assembly on the reference genome. Then, a possible splice site is recognized based on the adjacent exons. Additionally, the sequences not positioned on the reference genome are collected to establish the vacancy seed index. Finally, the vacancy expansion is compared in order to obtain the possible splice sites. According to a test reported by the authors, TopHat processed 2.2 million reads per hour, whereas QPALMA processed approximately 180,000. However, the performance will be poor when the depth of sequencing is low or the intron is very short because the algorithm adopts exon islands.

SpliceMap consists of four main steps: half-read mapping, seeding selection, site search, and paired-end filtering. First, SpliceMap splits the read into halves. Alignment positioning is performed between each portion and the gene sequence. Then, the remaining half is positioned on the downstream region within the range of the longest intron. This approach requires the read length to be at least 50 bp. Therefore, SpliceMap cannot process read lengths <50 bp.

When we compared SpliceMap with ERANGE, ERANGE discovered 160,899 sites, whereas SpliceMap accurately predicted 127,043 sites. Moreover, 24,274 of the 151,317 sites discovered by SpliceMap were not discovered by ERANGE, of which 23,020 represent new splice sites. However, these new sites are unconfirmed. The MapSplice software appeared after TopHat and SpliceMap. MapSplice is not based on the characteristics of splice sites or the length of an intron. It also has the potential to discover new sites and can adapt the length of the read.

The emergence of SOAPsplice improved the evaluation standard of splice site prediction software. SOAPsplice not only depends on the number of recognition splice sites but also emphasizes a high accuracy and low error rate. The experiment described in the next section revealed that the performance of SOAPsplice was comparatively outstanding. SplitSeek is strict with regard to the format of the input data and only supports data generated by ABI SOLiD. Moreover, because the input data are processed by a complete ABI transcriptome analysis tool, the application is not very wide. HMMSplicer is similar to SpliceMap but possesses several innovations. First, it divides the read into halves and compares halves with the genome sequence. The exon boundary (i.e., the $5'$ terminal) is obtained using the hidden Markov model (HMM). Second, the remaining half is positioned downstream the first half to determine the boundary $3'$ terminal of the intron. Both common (GT-AG, GC-AG, and AT-AC) and uncommon splice sites are recorded during this process. Finally, the scores of candidate loci are graded using the scoring algorithm.

3.3. Aligning Spliced Reads to the Reference Genome. Read lengths generated by all types of sequencing platforms are growing concomitant with the development of deep sequencing and RNA-seq technology. In the early days, read lengths were usually approximately 32 bp, and most of the software programs did not consider the location of the spliced reads on the reference genome. However, with the generation of longer reads, new requirements were put forward for locating software.

Reads mapping and alternative splicing detection are two steps in an analysis workflow. RNA read alignment is the precursor step and splice isoform detection is the successor step. Splice isoform detection tools include Cufflinks [81] and Scripture [82]. Cufflinks is a software tool for detecting the specific expression genes. If users have two groups of RNA-Seq data, such as ill and normal persons, it would be better to employ Cufflinks for the key genes detection. Scripture is a method for transcriptome reconstruction that relies solely on RNA-Seq reads and an assembled genome to build a transcriptome *ab initio*.

Researchers applied the preprepared splice site database when they began trying to align spliced reads to the reference genome. However, the existing annotation of the transcriptome was far from being perfect. Therefore, some researchers once again began using BLAT to locate reads.

The TopHat software program solved these problems and thus became widely used by researchers; moreover, its vision has been expanding in every release from its initial release. In addition to its ability to align spliced reads to the reference genome, TopHat can also predict possible splice sites. These splice sites play an important role in improving the annotation of the transcriptome. The initial vision of TopHat had many limitations; however, the adoption of new methods in the software updates has improved TopHat's performance.

With the development of sequencing technologies, reads with lengths >100 bp have been produced on a large scale. These reads may span one or more spliced sites, which introduces difficulty in aligning spliced reads. The SpliceMap software is capable of processing longer reads (read lengths > 50 bp). To process these long reads, SpliceMap divides the reads into overlapping short read fragments. Then, they are annotated with the locating information of whole reads based on the locating information of the short read fragments.

MapSplice is another package that aligns spliced reads to the reference genome, although it applies a different method. The MapSplice algorithm is suitable for all types of read lengths. It is similar to SpliceMap in that it does not use continuous aligning of the reads to create an exon library in advance. Because the MapSplice package does not depend on spliced read signal information when aligning reads, it can locate some reads that SpliceMap cannot align. It can also be used to predict new spliced reads with no spliced read signal information. Another advantage of the MapSplice package is its high efficiency compared with most other software.

Package SeqSaw was proposed by Wang et al. [83] and is totally different from TopHat and MapSplice. It was similar to the SpliceMap package in its early releases. However, SeqSaw use has dynamically changed to Hash Table to reduce the search space. The core algorithm of SeqSaw is focused on locating short reads to the genome. There are very few introns >400 Kb in the known mammalian genome. Thus, we can define intron lengths as being less than a certain value, with a default value of 400 Kb. Users can adjust the value according to the needs of different species or datasets. However, SeqSaw uses certain means and performs a large amount of optimization, which greatly reduces the search space.

The R package DEGseq [83] has been proposed to detect small changes in the genetic expression of each sample. It is used to assess the trend of background noise in MA due to technological repeats. Figure 2 shows the working process of DEGseq.

The difference between a DNA aligner and an RNA aligner is that an RNA aligner can tolerate extra-long deletions (introns) while DNA aligners cannot [84]. Moreover, many RNA aligners are constructed based on DNA aligners (i.e., TopHat is built based on Bowtie). STAR is the latest and most popular RNA-seq alignment tools. In addition to unbiased de novo detection of canonical junctions, STAR can discover noncanonical splices and chimeric (fusion) transcripts and is also capable of mapping full-length RNA sequences [85].

4. Experiments Using State-of-the-Art Software Tools

HMMSplicer, SOAPsplice, TopHat, and STAR were used to perform the following analysis of Illumina/Solexa output

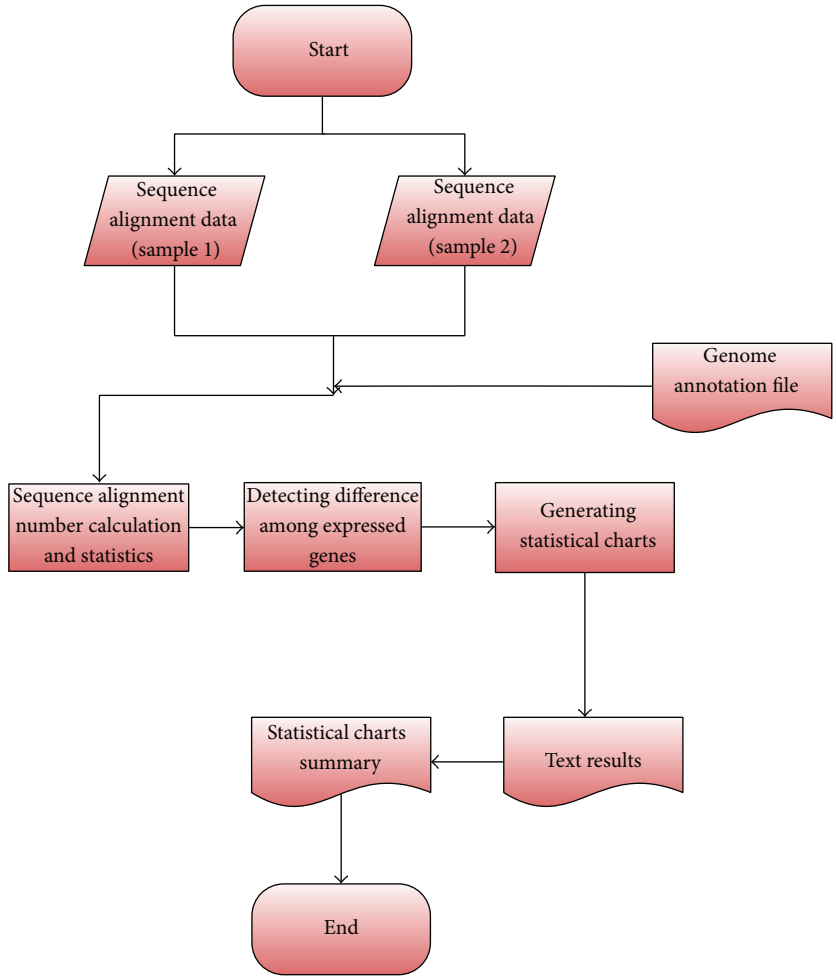


Figure 2: Working process of DEGseq.

data. The reference genome data are from the tenth human chromosome. The gene sequence was processed into RNA-seq sequences with different read lengths and different sequencing depths as the test data for SOAPsplice and TopHat. HMMSplicer does not support double-end sequencing data, so each pair of FASTQ data was merged into a FASTQ file as the test data for HMMSplicer.

Figure 3 shows that, in the premise of the 50 bp read length, each type of software predicts an increase in the number of loci that increases with the development of sequencing technologies. The accuracy of TopHat is poorer compared with the other two programs within a sequencing depth range of 1x to 10x, and the error rate is still high. The accuracy of TopHat increased rapidly after the sequencing was deepened. SOAPsplice and TopHat performed well in the aspect of accuracy, although the error rate was significantly worse for TopHat. STAR works best among the four tested tools. SOAPsplice and STAR performed well in both aspects.

5. Conclusion

In this study, we analyzed and compared the current AS-associated algorithms and software. We summarized the present situation of AS. The read mapping, including AS and site recognition algorithms, remained the focus of the current research. We aimed to improve the algorithm's quality in order to increase the number of prediction sites as much as possible and to meet the high-accuracy rate. RNA-seq data size is very large due to the continuous development of next-generation sequencing technology. This study represents a broad platform for AS and other fields of bioinformatics. This review of experimental and research methods for AS may be helpful for other researchers.

Although high-throughput sequencing has given rise to an unprecedented opportunity for the study of AS, few scholars study AS based on RNA-seq data. Therefore, the available algorithms and software are not rich compared with those based on EST/cDNA theory. Significant differences are found in the alignment step between the algorithms and the software using next-generation technology. This step represents the critical step based on the study of RNA-seq data. The software tools and algorithms need to be considered in parallel as the read data becomes more massive [86]. Genome-wide analysis will be the hot topic for all alternative and epigenetic research fields [87]. Moreover, many of the

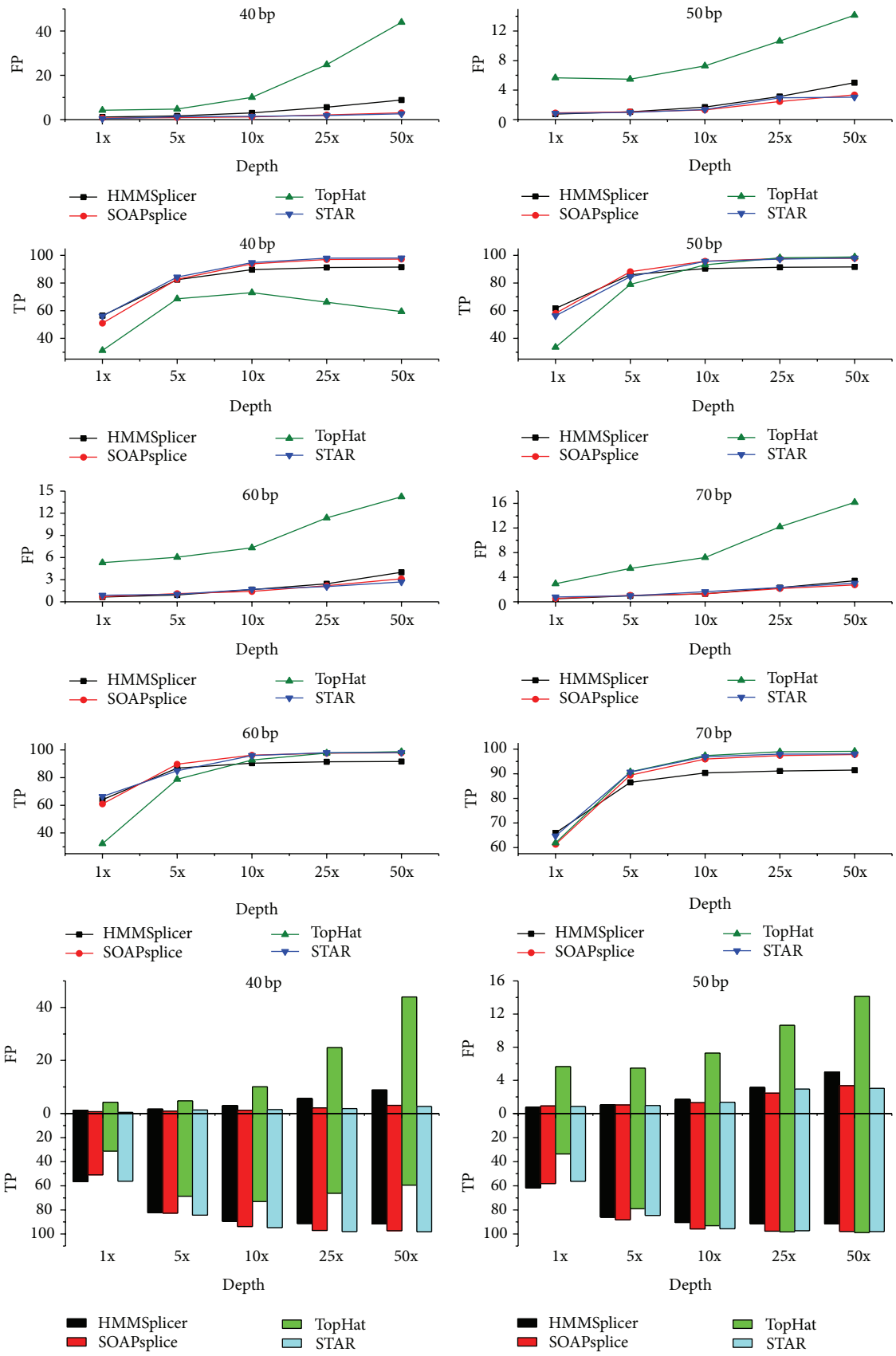


Figure 3: Continued.

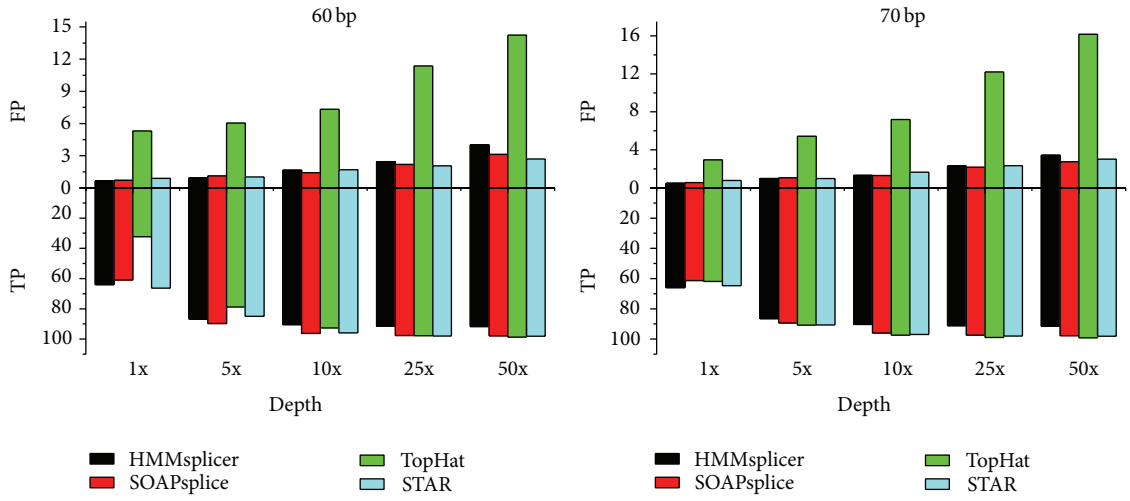


Figure 3: Comparison of HMMSplicer, SOAPsplice, STAR, and TopHat.

special databases based on RNA-seq data are not perfect. The corresponding new research methods and databases will be perfected with the constantly developing study of AS.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. M. Johnson, J. Castle, P. Garrett-Engele et al., "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays," *Science*, vol. 302, no. 5653, pp. 2141–2144, 2003.

[2] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.

[3] B. Modrek and C. Lee, "A genomic view of alternative splicing," *Nature Genetics*, vol. 30, no. 1, pp. 13–19, 2002.

[4] M. Dutertre, S. Vagner, and D. Auboeuf, "Alternative splicing and breast cancer," *RNA Biology*, vol. 7, no. 4, pp. 403–411, 2010.

[5] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

[6] S. Hua, W. Yun, Z. Zhiqiang, and Q. Zou, "A discussion of microRNAs in cancers," *Current Bioinformatics*, vol. 9, no. 5, pp. 453–462, 2014.

[7] S. Stamm, J.-J. Riethoven, V. Le Texier et al., "ASD: a bioinformatics resource on alternative splicing," *Nucleic Acids Research*, vol. 34, pp. D46–D55, 2006.

[8] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[9] B. Liu, J. Yi, A. Sv et al., "QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions," *BMC Genomics*, vol. 14, no. 8, article S3, 2013.

[10] L. T. Chow, R. E. Gelinas, T. R. Broker et al., "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA (Reprinted from Cell, vol 12, pg 1–12, 1977)," *Reviews in Medical Virology*, vol. 10, no. 6, pp. 362–369, 2000.

[11] M. T. Dimon, K. Sorber, and J. L. DeRisi, "HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data," *PLoS ONE*, vol. 5, no. 11, Article ID e13875, 2010.

[12] Q. Zou, X. Li, Y. Jiang, Y. Zhao, and G. Wang, "Binmempredict: a web server and software for predicting membrane protein types," *Current Proteomics*, vol. 10, no. 1, pp. 2–9, 2013.

[13] W. G. Fairbrother, R.-F. Yeh, P. A. Sharp, and C. B. Burge, "Predictive identification of exonic splicing enhancers in human genes," *Science*, vol. 297, no. 5583, pp. 1007–1013, 2002.

[14] D. L. Black, "Mechanisms of alternative pre-messenger RNA splicing," *Annual Review of Biochemistry*, vol. 72, pp. 291–336, 2003.

[15] P. Bonizzoni, R. Rizzi, and G. Pesole, "Computational methods for alternative splicing prediction," *Briefings in Functional Genomics and Proteomics*, vol. 5, no. 1, pp. 46–51, 2006.

[16] B. Modrek, A. Resch, C. Grasso, and C. Lee, "Genome-wide detection of alternative splicing in expressed sequences of human genes," *Nucleic Acids Research*, vol. 29, no. 13, pp. 2850–2859, 2001.

[17] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.

[18] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[19] L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller, "A computer program for aligning a cDNA sequence with a genomic DNA sequence," *Genome Research*, vol. 8, no. 9, pp. 967–974, 1998.

[20] N. Kim, S. Shin, and S. Lee, "ECgene: genome-based EST clustering and gene modeling for alternative splicing," *Genome Research*, vol. 15, no. 4, pp. 566–576, 2005.

[21] T. Castrignanò, R. Rizzi, I. G. Talamo et al., "ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization," *Nucleic Acids Research*, vol. 34, pp. W440–W443, 2006.

[22] S. J. Wheelan, D. M. Church, and J. M. Ostell, "Spidey: a tool for mRNA-to-genomic alignments," *Genome Research*, vol. 11, no. 11, pp. 1952–1957, 2001.

[23] J. Usuka, W. Zhu, and V. Brendel, "Optimal spliced alignment of homologous cDNA to a genomic DNA template," *Bioinformatics*, vol. 16, no. 3, pp. 203–211, 2000.

[24] T. D. Wu and C. K. Watanabe, "GMAP: a genomic mapping and alignment program for mRNA and EST sequences," *Bioinformatics*, vol. 21, no. 9, pp. 1859–1875, 2005.

[25] L. Wang, Y. Xi, J. Yu, L. Dong, L. Yen, and W. Li, "A statistical method for the detection of alternative splicing using RNA-seq," *PLoS ONE*, vol. 5, no. 1, Article ID e8529, 2010.

[26] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[27] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[28] B. Liu, J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, "Using distances between Top-n-gram and residue pairs for protein remote homology detection," *BMC Bioinformatics*, vol. 15, supplement 2, article S3, 2014.

[29] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[30] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identificationof human microRNAs by incorporatinga high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[31] Q. Zou, Y. Mao, L. Hu, Y. Wu, and Z. Ji, "miRClassify: an advanced web server for miRNA family classification and annotation," *Computers in Biology and Medicine*, vol. 45, no. 1, pp. 157–160, 2014.

[32] B. Liu, X. Wang, L. Lin, B. Tang, and Q. Dong, "Prediction of protein binding sites in protein structures using hidden Markov support vector machine," *BMC Bioinformatics*, vol. 10, article 381, 2009.

[33] Y. Zhang, B. Liu, Q. Dong, and V. X. Jin, "An improved profile-level domain linker propensity index for protein domain boundary prediction," *Protein and Peptide Letters*, vol. 18, no. 1, pp. 7–16, 2011.

[34] G. Wang, K. Qi, Y. Zhao et al., "Identification of regulatory regions of bidirectional genes in cervical cancer," *BMC Medical Genomics*, vol. 6, no. 1, article S5, 2013.

[35] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "nDNA-prot: identification of DNA-binding proteins based on unbalanced classification," *BMC Bioinformatics*, vol. 15, no. 1, article 298, 2014.

[36] B. Liu, J. Xu, X. Lan et al., "iDNA-Prot—dis: identifying dna-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS ONE*, vol. 9, no. 9, Article ID e106691, 2014.

[37] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[38] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[39] X.-Y. Cheng, W.-J. Huang, S.-C. Hu et al., "A global characterization and identification of multifunctional enzymes," *PLoS ONE*, vol. 7, no. 6, Article ID e38979, 2012.

[40] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[41] L. Cheng, Z.-G. Hou, Y. Lin, M. Tan, W. C. Zhang, and F.-X. Wu, "Recurrent neural network for non-smooth convex optimization problems with application to the identification of genetic regulatory networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 5, pp. 714–726, 2011.

[42] P. M. Feng, H. Lin, W. Chen, and Y. Zuo, "Predicting the types of j-proteins using clustered amino acids," *BioMed Research International*, vol. 2014, Article ID 935719, 8 pages, 2014.

[43] P. M. Feng, W. Chen, H. Lin, and K.-C. Chou, "IHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[44] W. Chen, P. Feng, and H. Lin, "Prediction of replication origins by calculating DNA structural properties," *FEBS Letters*, vol. 586, no. 6, pp. 934–938, 2012.

[45] W. C. Li, J. Z. Zhong, P. P. Zhu et al., "Sequence analysis of origins of replication in the *Saccharomyces cerevisiae* genomes," *Frontiers in Microbiology*, vol. 5, article 574, 2014.

[46] W. Chen, H. Lin, and P. M. Feng, "DNA physical parameters modulate nucleosome positioning in the *Saccharomyces cerevisiae* genome," *Current Bioinformatics*, vol. 9, no. 2, pp. 188–193, 2014.

[47] W. Chen, H. Lin, P.-M. Feng, C. Ding, Y.-C. Zuo, and K.-C. Chou, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[48] S.-H. Guo, E.-Z. Deng, L.-Q. Xu et al., "INuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

[49] P. M. Feng, W. Chen, and H. Lin, "Prediction of CpG island methylation status by intergrating DNA physicochemical properties," *Genomics*, vol. 104, no. 4, pp. 229–233, 2014.

[50] W. Chen, P.-M. Feng, E.-Z. Deng, H. Lin, and K.-C. Chou, "iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition," *Analytical Biochemistry*, vol. 462, pp. 76–83, 2014.

[51] H. Lin, E. Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Research*, vol. 42, no. 21, pp. 12961–12972, 2014.

[52] Z. Yu, H. Chen, J. You et al., "Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 4, pp. 727–740, 2014.

[53] Z. Yu, L. Li, J. You, H.-S. Wong, and G. Han, "SC^3: triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profiles," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1751–1765, 2012.

[54] W. Chen, H. Lin, P. Feng, and J. Wang, "Exon skipping event prediction based on histone modifications," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 241–249, 2014.

[55] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition," *BioMed Research International*, vol. 2014, Article ID 623149, 12 pages, 2014.

[56] Y. Q. Xing, L. R. Zhang, and L. F. Luo, "Prediction of alternative splicing sites of cassette exons and intron retention in human genome," *Acta Biophysica Sinica*, vol. 24, pp. 393–401, 2008.

[57] M. Wang and A. Marín, "Characterization and prediction of alternative splice sites," *Gene*, vol. 366, no. 2, pp. 219–227, 2006.

[58] A. Krogh, "Using database matches with HMMGene for automated gene detection in *Drosophila*," *Genome Research*, vol. 10, no. 4, pp. 523–528, 2000.

[59] S. Brunak, J. Engelbrecht, and S. Knudsen, "Prediction of human mRNA donor and acceptor sites from the DNA sequence," *Journal of Molecular Biology*, vol. 220, no. 1, pp. 49–65, 1991.

[60] S. M. Hebsgaard, P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouzé, and S. Brunak, "Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information," *Nucleic Acids Research*, vol. 24, no. 17, pp. 3439–3452, 1996.

[61] G. Parra, E. Blanco, and R. Guigó, "GeneID in *Drosophila*," *Genome Research*, vol. 10, no. 4, pp. 511–515, 2000.

[62] M. Pertea, X. Lin, and S. L. Salzberg, "GeneSplicer: a new computational method for splice site prediction," *Nucleic Acids Research*, vol. 29, no. 5, pp. 1185–1190, 2001.

[63] S. Degroeve, Y. Saeys, B. de Baets, P. Rouzé, and Y. van de Peer, "SpliceMachine: Predicting splice sites from high-dimensional local context representations," *Bioinformatics*, vol. 21, no. 8, pp. 1332–1338, 2005.

[64] Y. Xiaoling and S. S. Tang Tian, "Research progress and application of next-generation sequencing," *Biotechnology Bulletin*, vol. 10, pp. 76–80, 2010.

[65] K. F. Au, H. Jiang, L. Lin, Y. Xing, and W. H. Wong, "Detection of splice junctions from paired-end RNA-seq data by SpliceMap," *Nucleic Acids Research*, vol. 38, no. 14, pp. 4570–4578, 2010.

[66] X. Wang, X.-W. Wang, L.-K. Wang, Z.-X. Feng, and X.-G. Zhang, "A review on the processing and analysis of next-generation RNA-seq data," *Progress in Biochemistry and Biophysics*, vol. 37, no. 8, pp. 834–846, 2010.

[67] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[68] N. Homer, B. Merriman, and S. F. Nelson, "BFAST: an alignment tool for large scale genome resequencing," *PLoS ONE*, vol. 4, no. 11, Article ID e7767, 2009.

[69] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno, "SHRiMP: accurate mapping of short color-space reads," *PLoS Computational Biology*, vol. 5, no. 5, Article ID e1000386, 2009.

[70] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

[71] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[72] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[73] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[74] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Research*, vol. 38, no. 6, pp. 1767–1771, 2009.

[75] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing," *Genome Research*, vol. 17, no. 11, pp. 1697–1706, 2007.

[76] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt, "Assembling millions of short DNA sequences using SSAKE," *Bioinformatics*, vol. 23, no. 4, pp. 500–501, 2007.

[77] J. Butler, I. MacCallum, M. Kleber et al., "ALLPATHS: de novo assembly of whole-genome shotgun microreads," *Genome Research*, vol. 18, no. 5, pp. 810–820, 2008.

[78] F. de Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, "Optimal spliced alignments of short sequence reads," *Bioinformatics*, vol. 24, no. 16, pp. i174–i180, 2008.

[79] K. Wang, D. Singh, Z. Zeng et al., "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic acids research*, vol. 38, no. 18, article e178, 2010.

[80] A. Ameur, A. Wetterbom, L. Feuk, and U. Gyllensten, "Global and unbiased detection of splice junctions from RNA-seq data," *Genome Biology*, vol. 11, no. 3, article r34, 2010.

[81] C. Trapnell, A. Roberts, L. Goff et al., "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, no. 3, pp. 562–578, 2012.

[82] M. Guttman, M. Garber, J. Z. Levin et al., "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs," *Nature Biotechnology*, vol. 28, no. 5, pp. 503–510, 2010.

[83] L. Wang, X. Wang, X. Wang, Y. Liang, and X. Zhang, "Observations on novel splice junctions from RNA sequencing data," *Biochemical and Biophysical Research Communications*, vol. 409, no. 2, pp. 299–303, 2011.

[84] T. Steijger, J. F. Abril, P. G. Engström et al., "Assessment of transcript reconstruction methods for RNA-seq," *Nature Methods*, vol. 10, no. 12, pp. 1177–1184, 2013.

[85] A. Dobin, C. A. Davis, F. Schlesinger et al., "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[86] Q. Zou, X.-B. Li, W.-R. Jiang, Z.-Y. Lin, G.-L. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in Bioinformatics*, vol. 15, no. 4, pp. 637–647, 2014.

[87] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou, "An overview of SNP interactions in genome-wide association studies," *Briefings in Functional Genomics*, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 757530, 7 pages
http://dx.doi.org/10.1155/2015/757530

Research Article

Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites

Guohua Wang,[1,2] Fang Wang,[1] Qian Huang,[1] Yu Li,[2,3] Yunlong Liu,[4,5] and Yadong Wang[1]

[1] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[2] *Instrument Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[3] *School of Life Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[4] *Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[5] *Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*

Correspondence should be addressed to Guohua Wang; ghwang@hit.edu.cn and Yadong Wang; ydwang@hit.edu.cn

Received 5 December 2014; Accepted 16 April 2015

Academic Editor: Jennifer Wu

Transcription factors are proteins that bind to DNA sequences to regulate gene transcription. The transcription factor binding sites are short DNA sequences (5–20 bp long) specifically bound by one or more transcription factors. The identification of transcription factor binding sites and prediction of their function continue to be challenging problems in computational biology. In this study, by integrating the DNase I hypersensitive sites with known position weight matrices in the TRANSFAC database, the transcription factor binding sites in gene regulatory region are identified. Based on the global gene expression patterns in cervical cancer HeLaS3 cell and HelaS3-ifnα4h cell (interferon treatment on HeLaS3 cell for 4 hours), we present a model-based computational approach to predict a set of transcription factors that potentially cause such differential gene expression. Significantly, 6 out 10 predicted functional factors, including IRF, IRF-2, IRF-9, IRF-1 and IRF-3, ICSBP, belong to interferon regulatory factor family and upregulate the gene expression levels responding to the interferon treatment. Another factor, ISGF-3, is also a transcriptional activator induced by interferon alpha. Using the different transcription factor binding sites selected criteria, the prediction result of our model is consistent. Our model demonstrated the potential to computationally identify the functional transcription factors in gene regulation.

1. Introduction

In molecular biology and genetics, transcription factors (TFs) are proteins that bind to DNA sequences specifically, thereby regulating the transcription of genetic information from DNA to messenger RNA [1]. Once bound to DNA, these proteins can promote or block the recruitment of RNA polymerase to specific genes, making genes more or less active. Transcription factors are essential for the regulation of gene expression. Under the effect of transcription factors, the various cells of the body can function differently though they have the same genome. Transcription factors bind to one or more sequence sites, which are called transcription factor binding sites (TFBSs), attaching to specific DNA sequences of the genes they regulate [2]. Transcription factor binding sites can be defined as short DNA sequences (5–20 bp long) specifically bound by one or more transcription factors [3]. The transcription regulation is carried out by the interplay between transcription factors and their binding sites in DNA sequences; thus the prediction of TFBS is a vital step to understand the mechanism of transcription regulation and construct the network of transcription regulation. With the development of DNA microarrays and fast sequencing technique, many transcription factor binding sites have been identified by using experimental methods such as ChIP-chip and ChIP-Seq [4–6]. Because these methods will consume many experiment materials and many TFs have no corresponding antibodies, biological experimental methods cannot identify all TFs in the genome. Hence, many different computational methods have been proposed to search for additional members of a known transcription factor binding motif or discover novel transcription factor binding motifs.

In recent years, many computational methods such as regression based approaches have been proposed to discover transcription factor binding sites based on gene expression data. These methods can model the relationship between gene expression and transcription factor binding motifs in the promoter regions [7–9]. Bussemaker et al. proposed a simple linear model between gene expression and transcription factors using the TFBSs counts in the promoter region [10]. Based on this model, instead of the counts of TFBSs, Conlon et al. used position weight matrices (PWMs) to identify the motif candidates on upstream of genes [11]. In these previous methods, the whole promoter regions were always used as transcriptional regulatory regions that include TFBSs. As we all know, promoter regions are much longer than TFBSs; therefore, it will be better for TFBS prediction if we can narrow down the potential transcription factor binding region.

As early as the 1980s, the gene transcription was found to be related with the sensibility to DNase I (deoxyribonuclease I) of chromatin [12]. The sensibility to DNase I of chromatin which contains the actively transcribed genes is 100 times stronger than the one of the chromatin which does not contain the actively transcribed genes [13]. In 2013, Sheffield et al. [14] found that TFBSs were correlated with the DNase I hypersensitive (DHS) sites. The structure of the chromatin that contains DHS sites is looser, so that gene regulatory proteins can bind to these regions preferentially to exert biological functions [15–18]. Within the DHS sites, the regions are not digested easily and protected by specific proteins which probably are gene regulatory proteins such as transcription factors. In this study, the DHS sites were combined with gene expression data to deduce the target genes, and it was found that approximately 71 percent of DHS sites associated with at least one gene and some of these DHS sites associated with up to 44 genes, and among these genes the protein-coding genes were more than RNA genes. Using Encode ChIP-Seq data, the transcription factor binding sites were compared to the DHS sites, which showed highly overlapping percentage. Hence, the DHS sites in the promoter region can be used to identify TFBSs [19].

In our previous study, a model-based procedure has been developed to predict the functional TFBSs. The model utilized known position weight matrix to identify potential TFBSs in the gene promoter regions and built quantitative relationship between the TFBSs and gene expression levels. The transcriptional regulatory region was arbitrarily defined as the upstream region of transcription start site. In this study, we proposed a modified method that combined the DNase I hypersensitive sites with promoter regions to promote the accuracy of TFBS identification and recognize the regulatory function of transcription factors.

2. Methods

2.1. Biological Model System. The cervical cancer HeLaS3 cell, which is a clonal derivative of the parent HeLa cell, has been very useful in the clonal analysis of mammalian cell populations relating to chromosomal variation, cell nutrition, and plaque-forming ability. In recent years, as a tier of 2 cell types of ENCODE project, large sets of genome-wide study used the next generation sequencing technology to investigate gene expression, transcription factor binding sites, histone modification, and DNase I hypersensitive sites in HeLaS3 cell line. In this study, using genome-wide gene expression profile combined with DNase I hypersensitivity data, we developed a new method to predict the most important transcript factor in interferon alpha treated HeLaS3 cell line.

2.2. Gene Expression and DNase I Data Set. The gene expression profiles of HeLaS3 and HeLaS3 treated by interferon alpha for 4 hours were downloaded from Gene Expression Omnibus Database (GEO number: GSE15805), where Affymetrix Human Exon 1.0 ST Array was used to access the global gene expression patterns in 3 and 2 replicates. The DNase I data set of HeLaS3 used in this study was freely available for downloading from the uniform DNase I HS track of UCSC NCBI37/hg19 ENCODE (http://genome.ucsc.edu/encode/).

2.3. Differential Expressed Gene Identification. Each gene expression array of 3 HeLaS3 replicates and 2 HelaS3-ifnα4h replicates has been done the RMA normalization used Affymetrix Power Tools (APT) and removed the batch effects using ComBat in the previous study [20]. We utilized the Quantile Normalization [21] to eliminate the difference among the parallel experiments and then used the Scaling Normalization [22] to eliminate the difference between two cell types. The genes not reliably detected in at least one of the two cells were removed and only the protein-coding genes were picked up. After t-test calculation, we selected 197 probe sets by $P < 0.05$ and fold change $> \pm 2$; the expression levels of them were altered significantly. Removing the probe sets that were not reliably detected and that had absent annotation; finally, 181 differentially expressed genes [23] were left for analysis, in which 121 were upregulated and 60 were downregulated.

2.4. TFBS Prediction in DHS Sites. For the 181 differentially expressed genes, the DHS sites which located in the 1,000 bp upstream and 500 bp downstream of transcription start sites were picked up as transcriptional regulatory regions. Human RefSeq transcript annotation (hg19 genome assembly) and regulatory sequence were retrieved from the UCSC Genome Browser. 2188 position weight matrices (PWMs) in the TRANSFAC database were used to predict the transcription factor target genes. For each TF-DHS pair, the similarity scores were calculated by scanning the PWM of the transcription factor along the sequence of DHS site and the maximum score was selected as the binding affinity between the transcription factor and DHS site. For each PWM, we selected top 5000 DHS sites with highest similarity scores in genome-wide as potential TFBS.

2.5. The Prediction of Functional Transcription Factor. In order to describe the correlation between the genes expression levels and the binding affinity of transcription factors in

DHS sites, a simplified quantitative relationship is established using a linear model:

$$g_k = \sum_{i \in T_k} \left(\sum_m d\,[m,i] \right) x_i, \tag{1}$$

where g_k is the logarithmic ratio of mRNA expression levels of the kth gene in the treatment group comparing to control group, $d[m,i]$ is the matching score of ith PWM in the mth DHS sites within transcriptional regulatory region of the kth gene, T_k is the number of all the TFBSs having occurrences in the regulatory region of the kth gene, and x_i is the functional level of the ith PWM. The biological implication of this equation is that the measured gene expression level g_k is modeled by the effect of transcription, controlled by 5′ *cis*-acting elements. Because the expression level of genes we used in this study was Log2 RMA expression value, g_k was calculated according to the following formulation:

$$g_k = s_{k,\text{Treatment}} - s_{k,\text{Control}}, \tag{2}$$

where $S_{k,\text{Treatment}}$ is the logarithmic ratio of mRNA expression levels of the kth gene in the treatment group (HelaS3-ifnα4h) and $S_{k,\text{Control}}$ is the logarithmic ratio of mRNA expression levels of the kth gene in the control group (HelaS3).

The linear model only described the quantitative relationship between gene expression levels and PWMs of one differentially expressed gene. Thus, the model can be rewritten in a matrix formulation:

$$\begin{aligned} Z &= (CD)\,X, \\ X &= \left([CD]^T [CD] \right)^{-1} [CD]^T Z, \end{aligned} \tag{3}$$

where $Z = (g_k)$; $X = (x_i)$ and C is the marking matrix recording whether the DHS sites are within the transcriptional regulatory regions of differentially expressed genes or not. If the jth DHS site is within the transcriptional regulatory region of the ith gene, $C[i, j] = 1$; otherwise $C[i, j] = 0$. D is the score matrix representing the maximum score of each motif candidate in each DHS site. The model error based on a given selection of TFs will be defined as the sum square of the differences between observed and predicted mRNA expression levels:

$$e = \sum_{k=1}^{n} \left(g_k - \sum_{i \in T_k} \left(\sum_m d\,[m,i] \right) x_i \right)^2, \tag{4}$$

where e is the error of this model and n is the total number of differentially expressed genes. This equation can be rewritten in a matrix formulation:

$$\begin{aligned} \text{Err} &= \|Z - (CD)\,X\| \\ &= \left\| Z - (CD) \left([CD]^T [CD] \right)^{-1} [CD]^T Z \right\|. \end{aligned} \tag{5}$$

In this study, we iteratively computed the model error of each PWM for $N_p = 100{,}000{,}000$ times. In each iteration, the program selected $n_t = 5$ PWM candidates randomly. The model error of each set of PWMs was calculated. Meanwhile, we assigned a score value, transcription factor's contribution value (TFCV), for each PWM candidate. The TFCV can be calculated by the following formulation:

$$\text{TFCV} = \sum_N \frac{1}{\text{Err}^2}, \tag{6}$$

where Err is the model error and N is the number of selected PWM candidates in each iteration. If Err is smaller, namely, TFVC score is higher, the transcriptional function of PWM corresponding transcription factor will be more significant. Meanwhile, the cumulative TFs' functional levels (TFL) were calculated by the sum of x.

The program of functional transcription factor prediction can be summarized as follows.

(1) Calculate the matrix Z of expression levels of all the genes in the HelaS3-ifnα4h comparing to the HelaS3.

(2) Extract the DNA sequences of DHS sites of HelaS3 and calculate the score matrix D using PWM. For each PWM, the threshold value (ts) is set as the 5000th highest score.

(3) Construct the matrix C by comparing the position of DHS site and gene's regulatory region coordinate in the genome.

(4) Randomly pick n_t PWMs from all 2188 PWM candidates.

(5) Calculate the predicted model error Err.

(6) Calculate the TFCV and TFL of each PWM which is randomly picked in this iteration.

(7) Add the current transcriptional contribution score to the cumulative TFs' contribution value (TFCV) and add the current function level to the cumulative TFs' functional levels (TFL).

(8) Repeat the program (4–7) N_p times.

3. Results

3.1. Overlapping between DHS Sites and TFBS of HelaS3. The transcription factors ChIP-Seq data [16, 17] and DNase I hypersensitivity sites of HelaS3 cells were downloaded from the UCSC Genome Browser. After filtering out the ChIP-Seq experiments with poor quality, 42 TFBS profiles were considered the overlapping analysis with DHS sites in HelaS3 cells (Figure 1). Notably, we found that the binding sites of 26 transcription factors had more than 90% overlap and only 5 factors had less than 70% overlap with DHS sites. Among these 5 factors, CTCF which often acts as a chromatin "insulator" creates boundaries between topologically associating domains in chromosomes. Therefore, transcription factors tend to bind to the DHS sites and we can utilize the DHS sites to improve the accuracy of transcription factor binding sites prediction.

3.2. Functional Transcription Factor Identification. Potential PWMs which corresponded to the binding sequence of

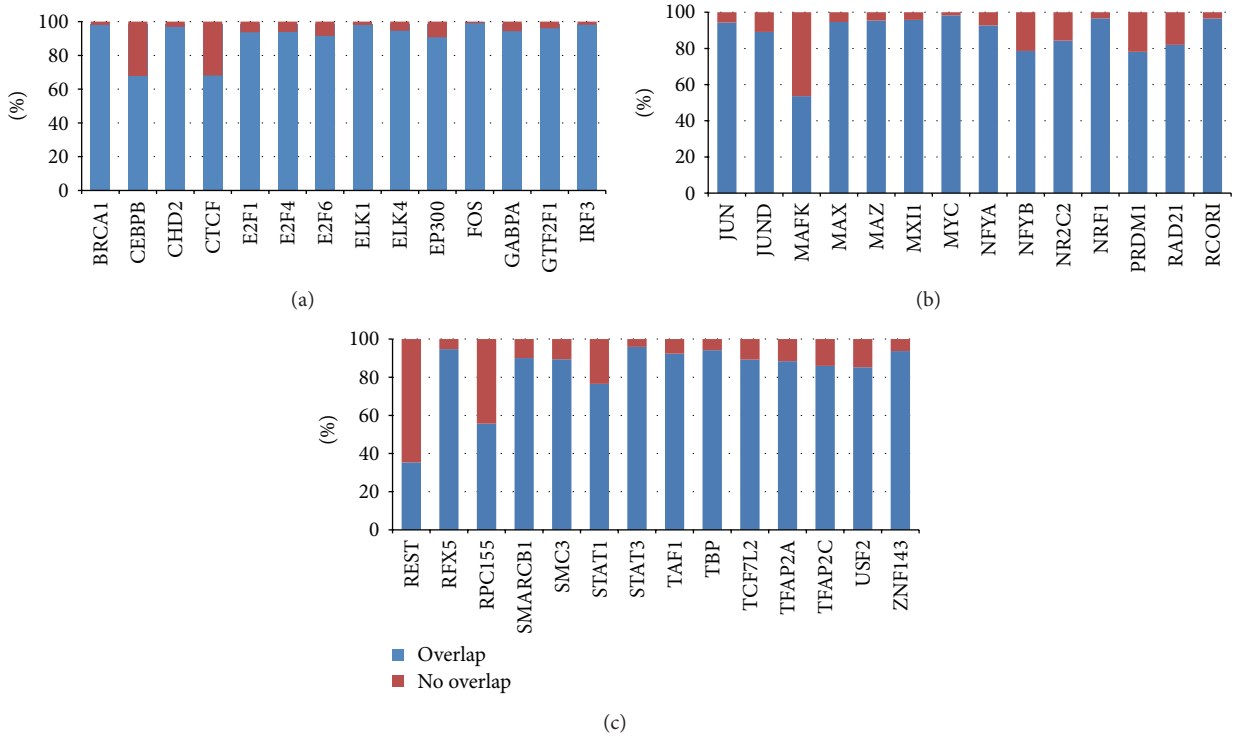


Figure 1: Overlapping between transcription factors binding regions and DHS sites. The blue bar and red bar represent the percentage of transcription factors that overlap and do not overlap with the DNase I hypersensitive sites, respectively.

a specific transcription factor were selected based on the binding affinity within DHS sites in the gene promoter region, as detailed in the methods. In order to predict the transcription factor binding sites, we calculated the score matrix D which stored the maximum scores as the binding affinity between the transcription factors and DHS sites. For each PWM, we selected top 5,000 matching positions with the highest similarity scores in the DHS sites genome-wide as potential TFBSs. After calculating our model iteratively, potential PWMs were selected based on the TFCVs of all PWM candidates. The histogram of TFCVs score of PWMs candidates is shown in Figure 2. In these PWM candidates, not all of them are real functional transcription factor binding sites. According to the methods, if the TFCV scores of PWMs are higher, their contributions to the alteration of gene expression are more significant. We selected the top 10 PWMs with the highest TFCV scores. The TFCV scores and the TFL values of these 10 PWM candidates are shown in Table 1. Significantly, 6 out 10 PWMs, including IRF, IRF-2, IRF-9, IRF-1, and IRF-3, ICSBP, belong to interferon regulatory factor family and upregulate the gene expression levels responding to the interferon treatment. ISGF-3 is also a transcriptional activator induced by interferon alpha. Among 10 PWMs, 9 received positive TFL values. This implies the increased capability of the 5′-end promoters in initiating transcription after treatment with interferon alpha.

3.3. Comparison of the Different TFBS Selection. To verify the accuracy of our model, we repeatedly run our model by changing the number of TFBSs to top 1000, 2000, 3000, or

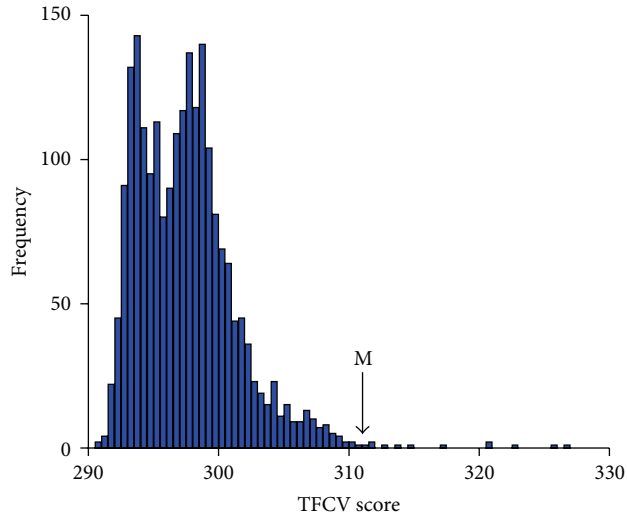


Figure 2: The histogram of TFCV scores for 2182 known PWMs. The x-axis is TFCV score and the y-axis is the frequency of the occurrence of TFCV for all known PWM.

4000 highest scores for each PWM. The TFCV profiles of each repeat computation are shown in Figure 3. We found that the distributions of TFCVs of all the PWM candidates in these 5 results were very similar. The Pearson correlation coefficient between the TFCV scores of each pair of predicted results was calculated. A heatmap corresponding to the Pearson correlation coefficient is shown in Figure 4. Obviously,

Table 1: Transcription factor's contribution value (TFCV) and estimated TFs' functional levels (TFL) of top 10 selected PWMs.

Index	ID	TF name	PWM description	TFCV	TFL
1	M00772	IRF	Interferon regulatory factor family	326.928	14830.189
2	M01882	IRF-2	Interferon regulatory factor 2	325.779	14680.555
3	M02771	IRF-9	Interferon regulatory factor 9	322.969	15127.858
4	M00258	ISGF-3	Interferon-stimulated response element	320.613	9914.496
5	M01881	IRF-1	Interferon regulatory factor 1	320.501	15363.707
6	M02767	IRF-3	Interferon regulatory factor 3	317.408	11305.011
7	M00699	ICSBP	Interferon consensus sequence-binding protein	314.717	7242.987
8	M00248	Oct-1	Octamer factor 1	313.642	6287.612
9	M01235	IPF1	Homeodomain-containing transactivator	310.253	6593.312
10	M01857	AP-2 alpha	Activating enhancer binding protein 2 alpha	309.403	−3725.557

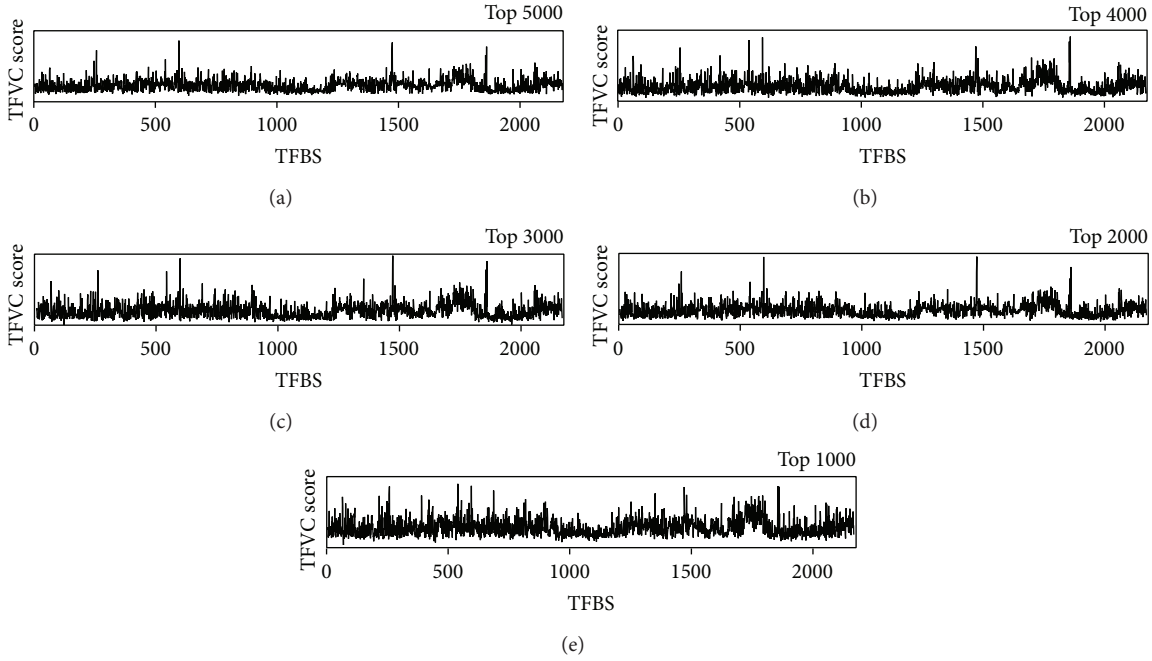


Figure 3: TFCV profile of 5 selected highest TFBS candidate models. The spectra of TFCV of all the PWMs while the threshold of potential TFBS is the 5000th, 4000th, 3000th, 2000th, or 1000th highest similarity score for each PWM. The x-axis corresponds to 2188 PWMs and the y-axis corresponds to TFCV scores.

the correlation between the prediction of top 1000 and top 5000 is the lowest (0.88), and the correlation between the prediction of top 4000 and top 5000 is the highest (0.96). The top 10 predicted PWMs with the highest TFCV score in all 5 calculations are shown in Table 2. Most of the top 10 PWMs are the same among these five prediction results, and most of them belong to interferon regulatory factor family.

Table 2: The top 10 transcription factors with the highest TFCV score in 5 selected highest TFBS candidate model.

Index	Top 1000	Top 2000	Top 3000	Top 4000	Top 5000
1	ICSBP	IRF-9	IRF-2	IRF-2	IRF
2	IRF	IRF	IRF	IRF	IRF-2
3	IRF-3	ICSBP	IRF-9	IRF-9	IRF-9
4	ISGF-3	IRF-3	IRF-1	ISGF-3	ISGF-3
5	IRF-9	IRF-2	IRF-3	IRF-1	IRF-1
6	IRF-1	ISGF-3	ISGF-3	IRF-3	IRF-3
7	IRF	IRF-1	ICSBP	ICSBP	ICSBP
8	EAR2	IRF-7	EAR2	Oct-1	Oct-1
9	IRF-5	IRF-1	IRF-1	IPF1	IPF1
10	RREB-1	EWSR1-FLI1	Lim1	AP-2 alpha	AP-2 alpha

4. Discussion

In this study, we modified the previous procedure Modif-Modeler to identify functional transcription factors. In the previous procedure, the transcription factor binding regions were set as the promoter regions [24]. To improve the accuracy of the identification of transcription factor binding sites, we reduced the searching space of transcription factor

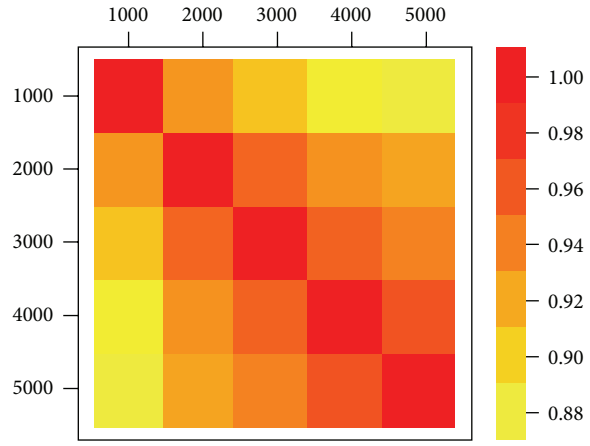


Figure 4: The cross-correlation coefficients of TFCV score among 5 selected highest TFBS candidate models.

binding regions. We have known that transcription factors tended to bind to DNase I hypersensitive sites; thus we combined the DNase I hypersensitive sites with promoter regions to construct a new model. In our model, using DHS sites within transcriptional regulatory region of each differentially expressed gene to replace all promoter regions, the binding regions of transcription factors were shortened and the accuracy of predicting transcription factor binding sites was improved. In this study, our model predicted some transcription factor binding sites whose functions differed as a result of interferon-α treatment.

Our modified model predicted that 9 of the top 10 transcription factors showed upregulatory effects on gene expression after interferon-α treatment which was clearly shown in Table 1. These predicted top 10 transcription factors with the largest TFCVs made significant contribution to the alteration of gene expression after interferon treatment. After being treated by interferon, some mechanisms of HelaS3-ifnα4h have changed compared with HelaS3 and some transcription factors responding to the interferon treatment have shown significant contribution to the alteration of gene expression. Obviously, most of the predicted TFs belong to interferon regulatory factor family, such as IRF-1, IRF-2, IRF-3, and IRF-9, ICSBP, and upregulate gene expression under interferon treatment [25–27]. Meanwhile a factor named interferon-stimulated response element (ISGF-3) also contributes to the alteration of gene expression significantly. It also indicates that our modified model can identify transcription factors which induced the gene expression change.

The identification of transcription factor binding sites is still a challenging and meaningful area. In the future, the identification of transcription factor binding sites will be very important and helpful for the understanding of the gene regulation mechanism [28]. Gene expression is regulated by many different elements synthetically. To predict different regulatory elements and understand their function, we also need to modify our model to adapt to various gene regulatory elements, such as microRNA and RNA binding proteins. In summary, focusing on the integration with DNase I hypersensitive sites allows high accuracy in our prediction procedure. As we all know, the identification of transcription factor binding sites can be used in clinic to find the change of regulatory elements in damaged or diseased cells and then help with the therapy of disease in the gene expression level [29]. We believe that our optimized method will contribute to an existing analytical network of gene expression.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contribution

Guohua Wang, Fang Wang, and Yadong Wang contributed to the design of the study. Guohua Wang, Yu Li, and Fang Wang designed and performed the computational modeling and drafted the paper. Qian Huang, Yu Li, and Yadong Wang participated in coordination, discussions related to result interpretation, and revision of the paper. All the authors read and approved the final paper.

Acknowledgments

This work was supported by grant from National High Technology Research and Development Program of China (2012AA020404), the National Natural Science Foundation of China (61371179), China Postdoctoral Science Foundation Funded Project (2012T50358, 20110491062, and 2014M551246), new century excellent talents support program from the Ministry of Education (NCET-13-0176), and the International Postdoctoral Exchange Fellowship Program 2013 (20130053).

References

[1] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional regulatory elements in the human genome," *Annual Review of Genomics and Human Genetics*, vol. 7, pp. 29–59, 2006.

[2] A. Jolma, J. Yan, T. Whitington et al., "DNA-binding specificities of human transcription factors," *Cell*, vol. 152, no. 1-2, pp. 327–339, 2013.

[3] G. Cuellar-Partida, F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, and T. L. Bailey, "Epigenetic priors for identifying active transcription factor binding sites," *Bioinformatics*, vol. 28, no. 1, pp. 56–62, 2012.

[4] B. Ren, F. Robert, J. J. Wyrick et al., "Genome-wide location and function of DNA binding proteins," *Science*, vol. 290, no. 5500, pp. 2306–2309, 2000.

[5] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park, "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *Nature Biotechnology*, vol. 26, no. 12, pp. 1351–1359, 2008.

[6] D. Park, Y. Lee, G. Bhupindersingh, and V. R. Iyer, "Widespread misinterpretable ChIP-seq bias in yeast," *PLoS ONE*, vol. 8, no. 12, Article ID e83506, 2013.

[7] G. Wang, X. Wang, Y. Wang et al., "Identification of transcription factor and microRNA binding sites in responsible to fetal alcohol syndrome," *BMC Genomics*, vol. 9, supplement 1, article S19, 2008.

[8] X. Dong, M. C. Greven, A. Kundaje et al., "Modeling gene expression using chromatin features in various cellular contexts," *Genome Biology*, vol. 13, no. 9, article R53, 2012.

[9] B. C. Foat, A. V. Morozov, and H. J. Bussemaker, "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE," *Bioinformatics*, vol. 22, no. 14, pp. e141–e149, 2006.

[10] H. J. Bussemaker, H. Li, and E. D. Siggia, "Regulatory element detection using correlation with expression," *Nature Genetics*, vol. 27, no. 2, pp. 167–171, 2001.

[11] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu, "Integrating regulatory motif discovery and genome-wide expression analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 6, pp. 3339–3344, 2003.

[12] Y. Kodama, S. Nagaya, A. Shinmyo, and K. Kato, "Mapping and characterization of DNase I hypersensitive sites in *Arabidopsis* chromatin," *Plant & Cell Physiology*, vol. 48, no. 3, pp. 459–470, 2007.

[13] A. P. Boyle, S. Davis, H. P. Shulha et al., "High-resolution map-ping and characterization of open chromatin across the genome," *Cell*, vol. 132, no. 2, pp. 311–322, 2008.

[14] N. C. Sheffield, R. E. Thurman, L. Song et al., "Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions," *Genome Research*, vol. 23, no. 5, pp. 777–788, 2013.

[15] R. Ciarpica, J. Rosati, and I. G. Cesaren, "Molecular recognition in helix-loop-helix leucine zipper domains," *The Journal of Biological Chemistry*, vol. 278, pp. 12182–12190, 2003.

[16] T. C. Gebuhr, G. I. Kovalev, S. Bultman, V. Godfrey, L. Su, and T. Magnuson, "The role of Brgl: a catalytic subunit of mammalian chromatin-remodeling complexes in T cell development," *Journal of Experimental Medicine*, vol. 198, no. 12, pp. 1937–1949, 2003.

[17] P. Blancafort, D. J. Segal, and C. F. Barbas III, "Designing transcription factor architectures for drug discovery," *Molecular Pharmacology*, vol. 66, no. 6, pp. 1361–1371, 2004.

[18] P. N. Cockerill, "Structure and function of active chromatin and DNase I hypersensitive sites," *The FEBS Journal*, vol. 278, no. 13, pp. 2182–2210, 2011.

[19] T. R. Mercer, S. L. Edwards, M. B. Clark et al., "DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements," *Nature Genetics*, vol. 45, no. 8, pp. 852–859, 2013.

[20] T.-Y. Chang, Y.-Y. Li, C.-H. Jen et al., "easyExon—a Java-based GUI tool for processing and visualization of Affymetrix exon array data," *BMC Bioinformatics*, vol. 9, article 432, 2008.

[21] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[22] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, article R25, 2010.

[23] A. Natarajan, G. G. Yardimci, N. C. Sheffield, G. E. Crawford, and U. Ohler, "Predicting cell-type-specific gene expression from regions of open chromatin," *Genome Research*, vol. 22, no. 9, pp. 1711–1722, 2012.

[24] J.-V. Turatsinze, M. Thomas-Chollier, M. Defrance, and J. van Helden, "Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules," *Nature Protocols*, vol. 3, no. 10, pp. 1578–1588, 2008.

[25] B. J. Barnes, J. Richards, M. Mancl, S. Hanash, L. Beretta, and P. M. Pitha, "Global and distinct targets of IRF-5 and IRF-7 during innate response to viral infection," *Journal of Biological Chemistry*, vol. 279, no. 43, pp. 45194–45207, 2004.

[26] W. Chen, S. S. Lam, H. Srinath et al., "Insights into interferon regulatory factor activation from the crystal structure of dimeric IRF5," *Nature Structural and Molecular Biology*, vol. 15, no. 11, pp. 1213–1220, 2008.

[27] T. Taniguchi and A. Takaoka, "The interferon-α/β system in antiviral responses: a multimodal machinery of gene regulation by the IRF family of transcription factors," *Current Opinion in Immunology*, vol. 14, no. 1, pp. 111–116, 2002.

[28] G. Gill, "Regulation of the initiation of eukaryotic transcription," *Essays in Biochemistry*, vol. 37, pp. 33–43, 2001.

[29] D.-J. Kleinjan and P. Coutinho, "Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease," *Briefings in Functional Genomics and Proteomics*, vol. 8, no. 4, pp. 317–332, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 979530, 17 pages
http://dx.doi.org/10.1155/2015/979530

Research Article

Active Microbial Communities Inhabit Sulphate-Methane Interphase in Deep Bedrock Fracture Fluids in Olkiluoto, Finland

Malin Bomberg,[1] Mari Nyyssönen,[1] Petteri Pitkänen,[2] Anne Lehtinen,[2] and Merja Itävaara[1]

[1] *VTT Technical Research Centre of Finland, P.O. Box 1000, 02044 Espoo, Finland*
[2] *Posiva Oy, Olkiluoto, 27160 Eurajoki, Finland*

Correspondence should be addressed to Malin Bomberg; malin.bomberg@vtt.fi

Received 6 January 2015; Accepted 4 March 2015

Academic Editor: Weixing Feng

Active microbial communities of deep crystalline bedrock fracture water were investigated from seven different boreholes in Olkiluoto (Western Finland) using bacterial and archaeal 16S rRNA, *dsr*B, and *mcr*A gene transcript targeted 454 pyrosequencing. Over a depth range of 296–798 m below ground surface the microbial communities changed according to depth, salinity gradient, and sulphate and methane concentrations. The highest bacterial diversity was observed in the sulphate-methane mixing zone (SMMZ) at 250–350 m depth, whereas archaeal diversity was highest in the lowest boundaries of the SMMZ. Sulphide-oxidizing ε-proteobacteria (*Sulfurimonas* sp.) dominated in the SMMZ and γ-proteobacteria (*Pseudomonas* spp.) below the SMMZ. The active archaeal communities consisted mostly of ANME-2D and Thermoplasmatales groups, although Methermicoccaceae, Methanobacteriaceae, and Thermoplasmatales (SAGMEG, TMG) were more common at 415–559 m depth. Typical indicator microorganisms for sulphate-methane transition zones in marine sediments, such as ANME-1 archaea, α-, β- and δ-proteobacteria, JS1, Actinomycetes, Planctomycetes, Chloroflexi, and MBGB Crenarchaeota were detected at specific depths. *Dsr*B genes were most numerous and most actively transcribed in the SMMZ while the *mcr*A gene concentration was highest in the deep methane rich groundwater. Our results demonstrate that active and highly diverse but sparse and stratified microbial communities inhabit the Fennoscandian deep bedrock ecosystems.

1. Introduction

Stable deep terrestrial subsurface locations are presently being considered for long-term geological disposal of spent nuclear fuel. Microbe-mediated processes may play a key role in the long-term stability and risk assessments of such storage. Dissolved sulphide produced by sulphate reducing bacteria (SRB), for example, may exert influence on spent nuclear fuel canister corrosion leading to mobility of radionuclides [1]. In Olkiluoto, Finland, spent nuclear fuel will be disposed approximately 450 m deep in the bedrock. Therefore, understanding the role and functionality of microbial communities in this environment is of critical importance for the safety of the spent nuclear fuel repository [2].

Deep subsurface microbial communities of the Fennoscandian Shield, including Olkiluoto, are functionally diverse and play a role in a variety of redox reactions, such as nitrate, iron, and sulphate reduction, as well as methanogenesis (e.g., [3–6]). While the presence of these processes has been confirmed by cultivation based techniques [4, 5, 7] and DNA-based PCR techniques [3, 6, 8], activity of these processes *in situ* remains uncertain.

In general, deep subsurface microbial communities appear to have extraordinarily low metabolic activity [6]. However, under certain environmental conditions, such as sulphate-methane transition zones (SMTZ), microbial activity appears to increase dramatically [9, 10]. At SMTZs in marine sediments, concentrations of H_2S increase (e.g., [11, 12])

possibly due to anaerobic oxidation of methane (AOM) and simultaneous reduction of SO_4^{2-}. In addition, both microbial cell concentration and microbial diversity have been seen to be elevated in sedimentary SMTZ environments [10]. Little is known of the activity, function, and composition of microbial communities in methane-rich deep terrestrial groundwater or terrestrial groundwater SMMZs.

Methane and sulphates are major constituents of Olkiluoto groundwater, residing in different groundwater layers [2]. Sulphate-rich water prevails at depths above 300 m below ground surface level (mbgsl) and methane-rich water dominates below 300 mbgsl. A sulphate-methane mixing zone (SMMZ) can be identified between 250 and 350 mbgsl [2]. In contrast to the clearly identifiable sharp SMTZs formed in anaerobic aquatic sediments [13, 14] the SMMZs in deep terrestrial groundwater are broad. In deep terrestrial subsurface, groundwater resides in bedrock fractures, which may be almost isolated and thereby exhibit stagnant groundwater or well connected with each other, which enables different degrees of groundwater flow. In addition, strong environmental changes, such as infiltration of surface water, crustal rebound, glaciation or deglaciation can affect the stability and position of the SMMZ [15].

Recently Pedersen et al. [16] simulated SMMZ mixing effect in Olkiluoto groundwater. By gradually increasing the concentration of sulphate in methane-rich and sulphate-poor groundwater over an experimental period of 103 days, the authors showed that the composition of the microbial community was strongly influenced by sulphate and methane. Several studies in Olkiluoto also show that the microbial communities in Olkiluoto groundwater are stratified and potentially affected by the groundwater SMMZ [3, 6, 17]. δ- and γ-proteobacteria are generally found in water layers above and in the SMMZ while β-proteobacteria become more abundant in the deeper methane-rich water [3, 6]. A clear increase in the number of methanogens was also detected simultaneously with a decrease in the number of sulphate reducing bacteria (SRB) in Olkiluoto deep groundwater [3, 17]. In addition, analysis of methyl coenzyme M reductase (*mcr*A) gene clone libraries demonstrates the presence of putative anaerobic methane oxidizing group 1 (ANME-1) archaea at 300–400 m depth [3].

Here, we extend this research and use RNA-targeted high-throughput (HTP) sequencing to investigate the active SRB and methanogen communities of the methane-rich deep groundwater around the depth of the nuclear waste repository rising up in to the SMMZ at the Olkiluoto site. In order to study the active microbial community in fracture water samples, the bacterial and archaeal 16S rRNA pools were also characterized and used as proxy for active (living) microbial cells. In addition, the abundance of SRB and methanogen communities was studied by qPCR targeting dissimilatory sulfite reductase (*dsr*B) and *mcr*A transcripts and genes.

2. Materials and Methods

2.1. Description of the Site. The island of Olkiluoto is the selected site for deep (approximately 450 mbgsl) geological disposal of spent nuclear fuel in Finland. The island has almost 60 boreholes drilled for research and monitoring purposes and studies on the chemistry and microbiology of the groundwater have been on-going since the 1980s [2]. The groundwater in Olkiluoto is stratified relative to physicochemical parameters [18]. From the surface to a depth of 30 mbgsl the water is of meteoric origin (i.e. precipitation) and the water type is fresh to brackish. The uppermost 100 mbgsl has a high concentration of dissolved inorganic carbon (as bicarbonates), and salinity (as total dissolved solids [TDS] and chlorine) increases with depth. Between 100 and 300 mbgsl, salinity is roughly similar to the present day Baltic Sea, but, below 300 mbgsl, the salinity increases up to 84 g TDS L^{-1} at 1000 mbgsl. Based on drill core logging, the bedrock of Olkiluoto consists mainly of gneiss (9% of the bedrock volume), migmatitic gneiss (64% of the bedrock volume), TGG (tonalite-granodiorite-granite) gneiss (8%), and pegmatitic granite (19%) [19]. In addition, of the migmatitic gneiss 67% is veined and 33% diatexitic gneiss.

Between 100 and 300 mbgsl, the SO_4^{2-} concentration is elevated in ancient (i.e., pre-Baltic) seawater derived groundwater. Below this layer, the methane concentration in the water increases and Cl^- dominates whereas SO_4^{2-} is almost absent. A mixing zone where methane-rich groundwater diffuses into sulphate-rich groundwater (a sulphate-methane mixing zone, SMMZ) can be identified at 250 to 350 mbgsl depth. This zone is characterized by increased concentration of sulphide and a decrease in sulphate and methane.

The temperature rises linearly with depth, from ca. 5-6°C at 50 mbgsl to ca. 20°C at 1000 mbgsl [20]. The pH of the water is slightly alkaline throughout the depth profile. Several aquifer zones, such as zones HZ20 or HZ21, span several different boreholes (Table 1).

2.2. Sampling. Deep groundwater samples (Table 1) from specific fracture zones were collected from seven different boreholes in Olkiluoto (Figure 1) between December 2009 and May 2010. Fracture zones were isolated by permanent or temporary inflatable packers as described previously [3]. Packer-sealed fracture zones were purged by pumping for at least four weeks prior to sampling in order to allow indigenous fracture water to fill the isolated borehole section. Anaerobic groundwater was pumped from the borehole directly in to an anaerobic chamber (MBRAUN, Germany) through a sterile, gas-tight polyacetate tube (8 mm outer diameter), where samples were collected in acid-washed, sterile 2 L Schott glass bottles (Duran Group GmBH, Germany). Microbial biomass for nucleic acid analyses was concentrated from 500 mL and 1000 mL samples by vacuum filtration through cellulose acetate membranes (0.2 μm pore size, Corning, MA, USA) inside the glove box. Filters were then cut out from the filter funnels and frozen on dry ice in sterile 50 mL cone tubes (Corning MA, USA). Frozen samples were transported on dry ice to the laboratory where they were stored at −80°C prior to analysis.

Samples for microbial cell counts were collected in acid-washed sterile, anaerobic 100 mL glass infusion flasks equipped with butyl rubber septa and aluminium crimp caps and transported to the laboratory at 4°C in a light-proof

Table 1: The geochemical and biological measurements from the samples collected from fracture fluids from seven different boreholes in Olkiluoto, Finland. The different boreholes are presented as sampling depths.

	296 m	328 m	347 m	415 m	559 m	572 m	798 m
Borehole	OL-KR13	OL-KR6	OL-KR23	OL-KR49	OL-KR2	OL-KR1	OL-KR29
Depth below ground surface (m)	−296.11	−328.37	−346.52	−415.45	−559.15	−572.24	−797.81
Water type	Brackish SO_4	Brackish SO_4	Saline	Saline	Saline	Saline	Saline
Transmissivity ($m^2\ s^{-1}$)	5.86×10^{-8}	1.31×10^{-7}	6.48×10^{-7}	4.37×10^{-7}	4.33×10^{-7}	5.50×10^{-7}	($<10^{-9}$)
Hydrogeological zone	HZ001		HZ20A		HZ21	HZ21	
Pump rate ($mL\ min^{-1}$)	22	104	20	172	23.9	62.1	6.1
Cumulative volume fracture fluid removed (L)	1129	5486	971	7509	1251	4492	496
Sampling date Microbiology	9.3.2010	18.5.2010	15.12.2009	14.12.2010	27.1.2010	26.1.2010	18.5.2010
Sampling date Chemistry	1.3.2010	10.5.2010	7.12.2009	1.12.2009	18.1.2010	18.1.2010	3.5.2010
Sampling date CH_4	6.3.2006	3.8.2005			18.3.2003	13.5.2003	4.4.2005
Temperature (°C)	19.6	11.6	17.6	11	14.8	12	17.7
pH	7.9	7.9	7.5	8.1	8.6	7.8	7.3
Ec ($mS\ m^{-1}$)	897	1832	2190	2670	4110	3770	7820
DIC ($mgC\ L^{-1}$)	27	4.1	3.9	<3	<3.75	<3.75	<21
NPOC ($mgC\ L^{-1}$)	10	<2.4	5.1	<3	11	5	<12
TDS ($mg\ L^{-1}$)	4994	10655	12733	15899	25459	23261	53205
Alk (m) $meq\ L^{-1}$	2.19	0.37	0.28	0.16	0.29	0.23	0.13
SO_4^{2-} ($mg\ L^{-1}$)	79.5	379	2.9	1.4	0.5	0.5	<2
S^{2-} ($mg\ L^{-1}$)	5.10	NA	0.62	0.02	<0.02	0.13	<0.02
NO_3 ($mg\ L^{-1}$)	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
NH_4 ($mg\ L^{-1}$)	0.07	0.03	<0.02	<0.02	<0.02	0.04	0.08
Fe^{2+} ($mg\ L^{-1}$)	<0.02	NA	0.08	0.53	<0.02	0.40	0.46
Na^{2+} ($mg\ L^{-1}$)	1320	2800	2530	3110	4980	4720	9150
K^+ ($mg\ L^{-1}$)	8.2	9.3	8.3	9.6	19	20	27
Ca^{2+} ($mg\ L^{-1}$)	460	1100	2100	2700	4600	3700	10000
Mg^{2+} ($mg\ L^{-1}$)	35	77	55	19	18	52	136
Cl^- ($mg\ L^{-1}$)	2920	6230	7930	9940	15700	14600	33500
CH_4 ($mL\ L^{-1}$)	22	22	NA	NA	386	272	920
TNC (mL^{-1})	4.2×10^5	1.0×10^5	2.5×10^5	1.5×10^4	5.9×10^4	8.7×10^4	2.3×10^4
*dsr*B gene copies mL^{-1}*	3.1×10^4 (8.6×10^3)	5.4×10^3 (2.2×10^3)	1.4×10^4 (6.8×10^3)	1.6×10^4 (9.9×10^3)	6.5×10^1 (2.0×10^1)	2.2×10^3 (2.9×10^2)	0
*dsr*B transcripts mL^{-1}*	1.4×10^2 (1.5×10^2)	1.2×10^2 (7.0×10^1)	2.9×10^2 (1.8×10^2)	3.7×10^0 (1.6×10^0)	0	2.0×10^1 (9.0×10^0)	0
*mcr*A copies mL^{-1}*	7.5×10^0 (2.5×10^0)	0	5.4×10^1 (2.7×10^1)	0	4.6×10^2 (5.2×10^0)	2.5×10^1 (4.8×10^0)	0
*mcr*A transcripts mL^{-1}*	0	0	0	0	0	0	0

NA: data not available.
*Figure in brackets shows standard error of mean (SEM).

container. The samples were analysed within 2 days of sampling.

2.3. Geochemistry. The geochemical data were provided by Posiva Oy and are presented in Table 1. Measurements were performed as described in Table 2.

2.4. Total Cell Counts. The total number of cells (TNC) was determined by fluorescent staining with 4′,6-diamidino-2-phenylindole (DAPI) [21] with slight modifications. A 5 mL subsample of each groundwater sample was stained with DAPI ($1\ \mu g\ mL^{-1}$) for 20 min at room temperature in the dark and collected on black polycarbonate Isopore Membrane

Figure 1: Map of Olkiluoto area where the different boreholes sampled in this study are indicated as open triangles. The arrows show the direction in which the boreholes lead. The scale bar is equal to 500 m.

filters (0.2 μm GTBP, Millipore, Ireland) with the Millipore 1225 Sampling Manifold (Millipore, USA) under low vacuum. The filters were rinsed with 1 mL filter sterilized 0.9% NaCl prior to and after filtration. Fluorescent cells were visualized under UV light with an epifluorescence microscope (Olympus BX60, Olympus Optical Ltd., Tokyo, Japan) and 1000x magnification. The number of cells was calculated from 30 random microscopy fields according to the magnification factor, filtered volume, and the surface area of the filter used [22].

2.5. Nucleic Acid Isolation. Microbial community nucleic acids (DNA and RNA) were isolated directly from the frozen cellulose-acetate filters with the PowerSoil DNA or PowerWater RNA extraction kit (MoBio Laboratories, Inc., Solana Beach, CA), respectively. Filters for DNA extraction were cut into 2 × 2 mm pieces with sterile scalpels in a laminar flow hood before insertion into the lysis tube. Nucleic acids were isolated according to the manufacturer's instructions except that for DNA extraction, the microbial cells were lysed by bead beating with a Precellys (Bertin Technologies, France) homogenizer for 30 s with 5 s increments at room temperature. The DNA and RNA from 500 mL and 1000 mL samples were eluted in 50 μL elution buffer and 100 μL elution buffer, respectively. Three replicate filters were used for DNA or RNA isolation. Negative isolation controls were performed from clean cellulose-acetate filter units in parallel with the samples using the same protocol and reagents as for the samples.

Residual DNA in the RNA extracts was checked by PCR with the primers used in this study (Table 3). If no PCR product was obtained, it was assumed that all residual DNA was successfully removed and the RNA extract was submitted to cDNA synthesis. If a PCR product was obtained, the RNA extract was treated with DNase (Promega, WI, USA) according to the manufacturer's instructions. cDNA was synthesized by first incubating 11.5 μL aliquots of RNA extract together with 250 ng random hexamers (Promega, WI, USA) and 0.83 mM final concentration dNTP (Finnzymes, Espoo, Finland) at 65°C for 5 minutes before cooling the reactions on ice for 1 minute. The reverse transcription was then performed with the Superscript III kit (Invitrogen), by adding 4 μL 5x First strand buffer, 40 U DTT, and 200 U Superscript III to the cooled reactions. To protect the RNA from degradation, 40 U of recombinant RNase inhibitor, RNaseOut (Promega, WI, USA), was used. The reactions were incubated at 25°C for 5 minutes, 50°C for 1 h, and 70°C for 15 min. Three parallel reactions were performed for each sample as well as for the reagent controls. The parallel reactions were subsequently pooled.

Table 2: Geochemical analysis methods and the detection limit of each assay used in this study. The data were obtained from Posiva Oy.

Parameter	Unit	Method	Detection limit
pH		pH meter, ISO-10532	
EC	(mS m^{-1})	Conductivity analyzer, SFS-EN-27888	5
NPOC	(mg L^{-1})	SFS-EN 1484	TC: 0.6 IC: 0.31 TOC: 0.3
TDS	(mg L^{-1})		
Alk	(meq L^{-1})	Titration with HCl	0.05
SO_4^{2-}	(mg L^{-1})	IC, conductivity detector	0.1
S^{2-}	(mg L^{-1})	Spectrophotometry	0.1
NO_3^-	(mg L^{-1})	FIA method, SFS-EN ISO11905-1	0.05
NH_4^+	(mg L^{-1})	Spectrophotometry, SFS 3032	
Fe^{2+}	(mg L^{-1})	Spectrophotometry	0.01
Na^{2+}	(mg L^{-1})	2007: FAAS, SFS3017, 3044 2008: ICP-OES	5 0.5
K^+	(mg L^{-1})	2007: FAAS, SFS3017, 3044 2008: ICP-OES	0.31 0.5
Ca^{2+}	(mg L^{-1})	2007: FAAS, SFS3017, 3044 2008: ICP-OES	0.02 0.1
Mg^{2+}	(mg L^{-1})	2007: FAAS, SFS3018 2008: ICP-OES	0.15 0.02
Cl^-	(mg L^{-1})	Titration	5
CH_4	(mL L^{-1} gas)	Gas chromatography	1 μL L^{-1} gas

Table 3: The primers used for amplification of different microbial groups for 454 pyrosequencing. The archaeal 16S rRNA and the *mcr*A gene transcripts were amplified using a nested PCR approach.

Target	Primer	Sequence	Fragment length (gene location)	Reference
Bacteria 16S rRNA	8F*	5′-AGAGTTTGATCCTGGCTCAG-3′	ca. 500 bp	[23]
	P2*	5′-ATTACCGCGGCTGCTGG-3′	(V1–V3)	[24]
Archaea 16S rRNA	A109f	5′-ACKGCTCAGTAACACGT-3′	ca. 800 bp	[25]
	Arch915R	5′-GTGCTCCCCCGCCAATTCCT-3′		[26]
	ARC344f*	5′-ACGGGGCGCAGCAGGCGCGA-3′	ca. 430 bp	[27]
	Ar744r*	5′-CCCGGGTATCTAATCC-3′	(V3-V4)	modified from [28]
Methanogens *mcr*A	*mcr*A412f	5′-GAAGTHACHCCNGAAACVATCA-3′	1.2 kb	[3]
	*mcr*1615r	5′-GGTGDCCNACGTTCATBGC-3′		[3]
	ME1*	5′-GCMATGCARATHGGWATGTC-3′	330 bp	[31]
	ME3r*	TGTGTGAAWCCKACDCCACC-3′		modified from [31]
Sulphate reducer *dsr*B	2060F*	5′-CAACATCGTYCAYACCCAGGG-3′	370 bp	[29]
	dsr4R*	5′-GTGTAGCAGTTACCGCA-3′		[30]

Primers marked with ∗ were equipped with adapter and barcode sequences at the 5′ ends, except if they were used for RT-qPCR. Primers marked with § were used in the qPCR without the adapters and barcodes.

2.6. Amplicon Library Preparation. Libraries for 454 high-throughput (HTP) amplicon sequencing were prepared by PCR from the cDNA samples. Bacterial 16S rRNA fragments covering the V1–V3 variable regions were amplified with primers 8F and P2 equipped with adapter and MID sequences at their 5′ end in a single round PCR (Table 3) [23, 24]. Archaeal 16S rRNA fragments were produced with a nested PCR using primers A109f and Arch915R [25, 26] for the first round and tagged primers ARC344f and Ar744r [27, 28] covering the V3-V4 variable areas for the second round. *Dsr*B fragments were amplified in a single round PCR with tagged primers 2060F [29] and *dsr*4R [30]. *Mcr*A fragments were obtained by nested PCR. Initially, a 1.2 kb *mcr*A fragment was amplified with primers *mcr*A412f and *mcr*1615r [3]. The product of this PCR was then amplified with tagged primers ME1 and ME3r modified from [31]. PCRs were performed with Phusion DNA polymerase (Finnzymes, Espoo, Finland) in 1x HF buffer. Each 50 μL reaction contained 0.5 mM dNTP and 1 μM of primers. The PCR conditions consisted of an initial denaturation step of 30 s at 98°C, followed by

35 cycles of 10 s at 98°C, 15 s at 55°C, 15 s at 72°C, and a final extension step at 72°C for 5 min. Two replicate samples were used for each borehole depth and a minimum of two amplification reactions were performed for each replicate sample, which were subsequently pooled prior to sequencing. All PCR reactions were also run with the negative nucleic acid extraction and reagent controls. The sequencing was performed at the Institute of Biotechnology, University of Helsinki, Finland, using the FLX 454 (454 Life Sciences, Branford, CT, USA).

2.7. Real-Time Quantitative PCR. The abundance of bacterial *dsr*B and archaeal *mcr*A genes and transcripts was determined by qPCR with KAPA SYBR Fast 2x Master mix for Roche LightCycler 480 (Kapa Biosystems, Inc., Boston, MA, USA). Reactions were performed in triplicate for each sample. Each reaction contained 1 μL of extracted DNA or cDNA as template and 5 pmol of both forward and reverse primers (Table 3). The qPCR was performed on a Roche LightCycler 480 (Roche Applied Science, Germany) on white 96-well plates (Roche Applied Science, Germany) sealed with transparent adhesive seals (4titude, UK). The qPCR conditions consisted of an initial denaturation at 95°C for 10 minutes followed by 45 amplification cycles of 15 seconds at 95°C, 30 seconds at 55°C, and 30 seconds at 72°C with a quantification measurement at the end of each elongation. A final extension step of three minutes at 72°C was performed prior to a melting curve analysis. The melting curve analysis consisted of a denaturation step for 10 seconds at 95°C followed by an annealing step at 65°C for one minute prior to a gradual temperature rise to 95°C at a rate of $0.11^{\circ}\mathrm{C\,s^{-1}}$ during which the fluorescence was continuously measured. The number of gene and transcript copies was calculated by comparing the amplification result (Cp) to that of a dilution series of plasmids containing *mcr*A or *dsr*B genes ranging from 0 to 10^7 gene copies per reaction as described in Nyyssönen et al. [3]. The lowest detectable standard concentration for the *dsr*B qPCR was 16 *dsr*B gene copies/reaction. In the *mcr*A qPCR assay, the lowest detectable standard had 100 *mcr*A copies/reaction. Template inhibition of the qPCR was tested by adding 2.17×10^4 plasmid copies containing fragment of the morphine-specific Fab gene from *Mus musculus* gene to reactions containing template DNA or cDNA and comparing the result to a dilution series of the plasmid as described in [3]. The inhibition of the qPCR assay by the template DNA was found to be low. The average Crossing point (Cp) value for the standard sample (2.17×10^4 copies) was 28.7 (±0.4 std), while for the DNA samples the Cp was 28.65–28.91 (±0.03–0.28 std) and for the cDNA samples was 28.69–28.96 (±0.02–0.23 std). Nucleic acid extraction and reagent controls were run in all qPCRs in parallel with the samples. Amplification in these controls was never higher than the background obtained from the no template controls.

2.8. Sequence Processing and Analysis. Sequence reads were trimmed with Mothur (v 1.31.2) [32] to remove adapter, barcode, and primer sequences and to exclude sequences that did not meet the quality criteria (i.e., no barcode and primer mismatches, no ambiguous nucleotides, maximum eight nucleotide long homopolymer stretches, and defined minimum length). The minimum length was 300 bp for bacterial 16S rRNA and *dsr*B sequences and 200 bp for archaeal 16S rRNA and *mcr*A sequences. The bacterial and archaeal 16S rRNA sequences were aligned with Mothur [32] using a Silva reference alignment [33] for bacterial (14 956 sequences) and archaeal (2 297 sequences) 16S rRNA gene sequences, respectively. The *dsr*B sequences were aligned with Geneious Pro (v 5.6, Biomatters Ltd., New Zealand) using a *dsr*AB model alignment [34] (97 sequences). The *mcr*A sequences were aligned with Mothur using a *mcr*A gene sequence model alignment (this study) (213 sequences). The alignments from the amplicon libraries were checked and manually corrected with Geneious Pro before further analysis with Mothur.

The sequences were divided into operational taxonomic units (OTUs) based on 97% sequence homology for the bacterial and archaeal 16S rRNA sequences and the *dsr*B sequences and 99% for the *mcr*A sequences. The sequencing coverage was evaluated by rarefaction analysis and the estimated species richness and diversity indices were calculated in Mothur.

The bacterial and archaeal 16S rRNA sequences were taxonomically classified with Mothur using the GreenGenes 13_8 database [35]. The representative sequences of the *dsr*B and *mcr*A OTUs were analysed using the Geneious Pro (Biomatters Inc., New Zealand). The *dsr*B and *mcr*A sequences were imported into Geneious Pro and aligned to reference sequences and most closely matching sequences determined against the NCBI database with blastn tool in Geneious Pro. The alignments were performed with Muscle [36] using default settings and the alignments were edited manually. The *mcr*A and *dsr*B sequences were subsequently translated to amino acid sequences before phylogenetic analyses. Phylogenetic analyses were performed on the alignments using PhyML [37] with the Jukes-Cantor (JC69) [38] substitution model for nucleic acid sequences and the Whelan-Goldman substitution model [39] for amino acid sequences. Bootstrap support for nodes was calculated based on 1000 random repeats.

For comparable α- and β-diversity analyses the data sets were normalized by random subsampling according to the sample with the lowest number of sequence reads, that is, 1200, 893, 2249, and 2324 sequences for archaea, bacteria, *dsr*B, and *mcr*A, respectively.

The sequences have been submitted to the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/) under accession numbers ERS514153–ERS514176.

2.9. Statistical Analyses. Statistical analyses were calculated with PAST v. 3.0 [40] in order to determine which of these parameters correlated most strongly with the detected taxa. The Shapiro-Wilk test [41] and Anderson-Darling test [42] were performed to analyze the normal distribution of the geochemical parameters. For sample parameters with $P < 0.05$ normal distribution were rejected and these parameters were excluded from the correlation calculations. The excluded parameters were DIC, bicarbonate, alkalinity, sulphate, S_{tot}, N_{tot}, Fe(II), F_{tot}, Sr, 16S rRNA gene copies

mL^{-1}, and *dsr*B transcripts mL^{-1} and *mcr*A genes mL^{-1}. Pearson's linear *r* correlation between presence and absence of different taxa in correlation to the geochemical parameters was calculated with PAST.

3. Results and Discussion

The crystalline bedrock of Olkiluoto has been chosen to host the deep geological repository for spent nuclear fuel in Finland. The spent nuclear fuel will be stored in copper canisters with nodular cast iron insert at 450 m depth and isolated from the bedrock by bentonite clay. Groundwater salinity and carbon content at different depths as well as the increase in the amount of CH_4 and H_2S and decrease in the amount of SO_4^{2-} at specific depths suggest the existence of a broad sulphate-methane mixing zone (SMMZ) in the groundwater at approximately 250–350 mbgsl depth [2]. At corresponding sulphate-methane transition zones (SMTZ) in marine sediments both the microbial activity and the diversity of the microbial communities increase dramatically [9, 43]. If the same kind of intensified activity occurs in groundwater SMMZs an increased risk may arise for, for example, microbially induced sulphate reduction aided corrosion of the waste capsules, release of radioactive waste, and mobilization of radionuclides.

In this study, we investigated the transcriptionally active microbial communities of the deep methane-rich groundwater spanning the depth of the future spent nuclear fuel repository. Triplicate groundwater samples from depths between 296 and 798 mbgsl from seven different boreholes in Olkiluoto were collected in order to characterize the active microbial communities around the depth of the planned repository (Table 1, Figure 1). The samples represented brackish SO_4^{2-} rich water and saline methane-rich water (as classified in [2]). The carbonate content in the groundwater generally decreased with depth whereas in deeper water the concentration of methane increased from almost none at 296 m to more than 900 mL L^{-1} gas at 800 mbgsl. The concentration of SO_4^{2-} was highest (379 mg L^{-1} groundwater) in the sample from 328 mbgsl and decreased radically with depth. The H_2S concentration was also highest at 296–347 mbgsl and decreased with depth.

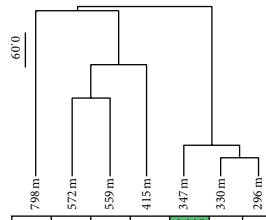
The TNC mL^{-1} groundwater varied between 4.2×10^5 mL^{-1} at 296 m and 1.5×10^4 mL^{-1} at 415 mbgsl with a general decline with depth (Table 1). HTP sequencing of bacterial and archaeal 16S rRNA with 454 technologies identified a total of 95 bacterial families and 27 archaeal families in the seven analyzed samples (Figures 2 and 3). The rarefaction analyses showed that the bacterial and archaeal communities were well characterized from 415 to 572 mbgsl (Figure 4). In the remaining samples, between 16 and 52% of the estimated bacterial and archaeal OTU richness was captured by sequencing.

*dsr*B gene transcripts were obtained from sequencing from depths between 296 mbgsl and 572 mbgsl, but not from the deepest sample from 798 mbgsl. The *dsr*B sequences belonged to six different SRB families and 14 genera (Figures 5 and 6). The *dsr*B transcript diversity was well covered showing between 81 and 98% of the estimated Chao1 OTU richness obtained. Transcripts of the *mcr*A genes were obtained for 454 sequencing with nested PCR amplification from four different depths, 328 m, 347 m, 572 m, and 798 mbgsl (Figure 7). The *mcr*A transcripts belonged to four methanogenic genera (Figure 8) that covered the Chao1 estimation of the total *mcr*A diversity.

Diversity of the active microbial communities was highest at sampling depths between 296 and 347 mbgsl, that is, in the SMMZ. At this depth, both bacterial diversity ($H' = 1.8$, normalized to equal number of sequence reads/sample) and SRB ($H' = 2.29$ and 2.65) diversity were the highest (Table 4). The highest archaeal diversity ($H' = 1.91$), in contrast, was seen in the lowest boundaries of the SMMZ at 347 mbgsl. The diversity of the methanogenic communities was low in all samples from which sequences were obtained by nested PCR ($H' = 0.42$–0.76).

3.1. Sulphate-Methane Mixing Zone (SMMZ). The structure of the active bacterial communities was similar between samples derived from similar depth of the different boreholes but changed with greater depth intervals (Figure 2). Sampling depths between 296 and 347 mbgsl contain the most H_2S and SO_4^{2-} rich water in this study and are influenced by a fraction of the methane-rich groundwater from deeper groundwater layers. Here, the most abundant bacterial group was ε-proteobacteria of the Helicobacteraceae family mostly belonging to the *Sulfurimonas*. This group formed 54–95% of the active bacterial communities as determined by the total number of sequences. ε-proteobacteria are believed to be enriched in the vicinity of SMTZs in marine sediments [44] and many are mesophilic, H_2- and sulphur-oxidizing chemolithoautotrophs [44–46]. They may play a profound role in recycling H_2S to SO_4^{2-} and are also a significant group in SMMZ microbial communities [10] where they fix CO_2 at the expense of sulphides and other electron donors. By fixing CO_2, they may account for a significant amount of assimilated carbon compounds available to microbial communities in deep subsurface environments [47]. The second largest group at 296–347 mbgsl was Desulfobacterales δ-proteobacteria forming 2–29% of the active community based on 16S rRNA (Figure 2). This is in accordance with the detection of the *dsr*B gene transcripts similar to uncultured group 1 Desulfobulbaceae of the Desulfobacterales family at this depth. These *dsr*B transcripts formed more than 69% of the *dsr*B transcripts at 296 mbgsl and showed a positive and significant correlation (>0.8, $P < 0.01$) with pH between 7.9 and 8.1. At 328 mbgsl, *dsr*B transcripts of the genera *Desulfotignum* and undefined *Desulfosarcina* of the Desulfobacteraceae were the most common. The amount of *dsr*B genes varied between 0.5 and 3.1×10^4 copies mL^{-1} at 296–374 mbgsl. In addition, the highest transcriptional activity of the *dsr*B genes, 1.2–2.9×10^2 transcripts mL^{-1}, was detected here, coinciding with the highest sulphate and sulphide concentrations and the lowest methane concentrations measured in this study.

At 296–347 mbgsl, a minor portion of the bacterial community belonged to methylotrophic β-proteobacteria and Verrucomicrobia, which may be capable of methane oxidation in the SMMZ (Figure 2). However, a more likely scenario



798 m	572 m	559 m	415 m	347 m	330 m	296 m				
				0.095			Unclass. Bacteria			
0.095		0.284					C111	Acidimicrobiales	Actinobacteria	Actinobacteria
		0.095					Unclass. Actinomyvtales	Actinomycetales		
		0.379					ACK-M1			
				0.095			Brevibacteriaceae			
3.602							Corynebacteriaceae			
0.095							Dermacoccaceae			
0.095							Frankiaceae			
				0.095			Kineosporiaceae			
2.749	0.947	16.004	0.473		0.095		Microbacteriaceae			
18.294		0.189		0.189			Micrococcaceae			
0.095		4.830	0.189		0.095		Nocardiaceae			
2.844		0.095				0.095	Propionibacteriaceae			
		3.409					OPB41			
		0.189					Armatimonadaceae	Armatimonadales	Armatimonadia	Armatimonadetes
0.095							Porphyromonadaceae	Bacteroidales	Bacteroidia	Bacteroidetes
0.190							Paraprevotellaceae			
				0.189		0.095	Unclass. Bacteroidetes			
						0.095	Cytophagaceae	Cytophagales	Cytophagia	
0.095	0.095	0.379					Weeksellaceae	Flavobacteriales	Flavobacteriia	
	0.095	0.095					Chitinophagaceae	Saprospirales	Saprospirae	
		0.095					BHI80-139			
						0.095	Caldilineaceae	Caldilineales	Anaerolineae	Chloroflexi
0.095							Roseiflexales		Chloroflexi	
		0.189		0.379		0.189	Dehalococcoidaceae	Dehalococcoidales	Dehalococcoidetes	
		0.284					SL56			
			1.231				Streptophyta		Chloroplast	Cyanobacteria
0.095							ML635J-21			
		0.379					Nostocaceae	Nostocales	Nostocophycideae	
0.190		1.042					Rivulariaceae	Stigonematales		
				0.284			Elusimicrobia			
		0.095					FBP			
46.256				0.095			Staphylococcaceae	Bacillales	Bacilli	Firmicutes
0.095							Planococcaceae			
0.095							Gemellaceae	Gemellales		
0.095							Aerococcaceae	Lactobacillales		
0.758							Carnobacteriaceae			
0.284							Enterococcaceae			
0.190							Lactobacillaceae			
2.464		0.095					Streptococcaceae			
		0.095					Unclassified Clostridiales	Clostridiales	Clostridia	
			0.095	0.095			Acidaminobacteraceae			
0.190			0.095				Clostridiaceae			
		0.095	0.189				Eubacteriaceae			
		3.504					Peptococcaceae			
3.981		0.284					Tissierellaceae			
		1.420					Desulforudales			
						0.095	Natranaerobiales			
		0.473					Erysipelotrichaceae	Erysipelotrichales	Erysipelotrichi	
0.095							Fusobacteriaceae	Fusobacteriales	Fusobacteria	Fusobacteria
						0.189	Victivallaceae	Victivallales	Lentisphaeria	Lentisphaerae
		0.095					ABY1			OD1
				0.095		0.095	OP3			
		0.189		0.095			SB-45		JS1	OP9
						0.095	SHA-43		Phycisphaerae	Planctomycetes
0.379	0.284	1.894		0.095	0.189	0.189	Caulobacteraceae	Caulobacterales	Alphaproteobacteria	Proteobacteria
0.190	0.095	2.557				0.095	Bradyrhizobiaceae	Rhizobiales		
					0.189		Hyphomicrobiaceae			
0.190		0.189					Methylobacteriaceae			
0.095							Rhizobiaceae			
0.190							Rhodobiaceae			
		1.989					Rhodobacteraceae	Rhodobacterales		
0.569							Acetobacteraceae	Rhodospirillales		
0.190		0.095	0.284			0.095	mitochondria	Rickettsiales		
		0.095					Unclass. Sphingomonadales	Sphingomonadales		
	0.095						Erythrobacteraceae			
1.422	3.504	0.947	0.758	0.095	0.095	0.379	Sphingomonadaceae			
					0.095		Unclass. Burkholderiales	Burkholderiales	Betaproteobacteria	
0.095	0.568	14.773	3.409	0.095	0.284	0.284	Comamonadaceae			
1.043							Oxalobacteraceae			
					0.095		Methylophilales			
0.095							Neisseriaceae	Neisseriales		
			1.799		0.095	1.894	Rhodocyclaceae	Rhodocyclales		
	0.095				0.379		Unclass. Deltaproteobacteria		Deltaproteobacteria	
				8.617	1.420	4.545	Desulfobulbaceae	Desulfobacterales		
0.284				20.170	0.947	0.473	Desulfobacteraceae			
			0.852				Desulfovibrionaceae	Desulfovibrionales		
		0.379					Desulfomicrobiaceae			
				0.189			Pelobacteraceae	Desulfuromonadales		
0.095							Cystobacterineae	Myxococcales		
				0.189	0.473	0.095	Syntrophaceae	Syntrophobacterales		
				0.095			Campylobacteraceae	Campylobacterales	Epsilonprot.	
0.569				54.545	94.60	86.17	Helicobacteraceae			
1.611							Enterobacteriaceae	Enterobacteriales	Gammaprot.	
0.095							Pasteurellaceae	Pasteurellales		
1.706							Moraxellaceae	Pseudomonadales		
6.635	94.223	41.572	90.625	8.996	0.947	1.136	Pseudomonadaceae			
		0.095		0.663		0.095	noFP_H7		AB16	SAR406
				3.598			MVP-15			Spirochaetes
				0.758		3.409	PL-11B10		MVP-15	
						0.095	Spirochaetaceae	Spirochaetales	Spirochaetes	
		0.663		0.095			Acholeplasmataceae	Acholeplasmatales	Mollicutes	Tenericutes
1.327							Deinococcus			Thermi
0.095		0.473					EW055		TM7-3	TM7
				0.095			Verrucomicrobia			

Figure 2: The relative distribution of bacterial 16S rRNA sequence reads belonging to specific bacterial families. The relative abundance of sequence reads are highlighted by color, where green represents the lowest relative abundance, yellow represents medium abundance, and red represents high abundance. The samples were clustered using the Morisita-Horn algorithm in Mothur. The data were normalized between the different samples to include 893 random sequence reads from each sample.

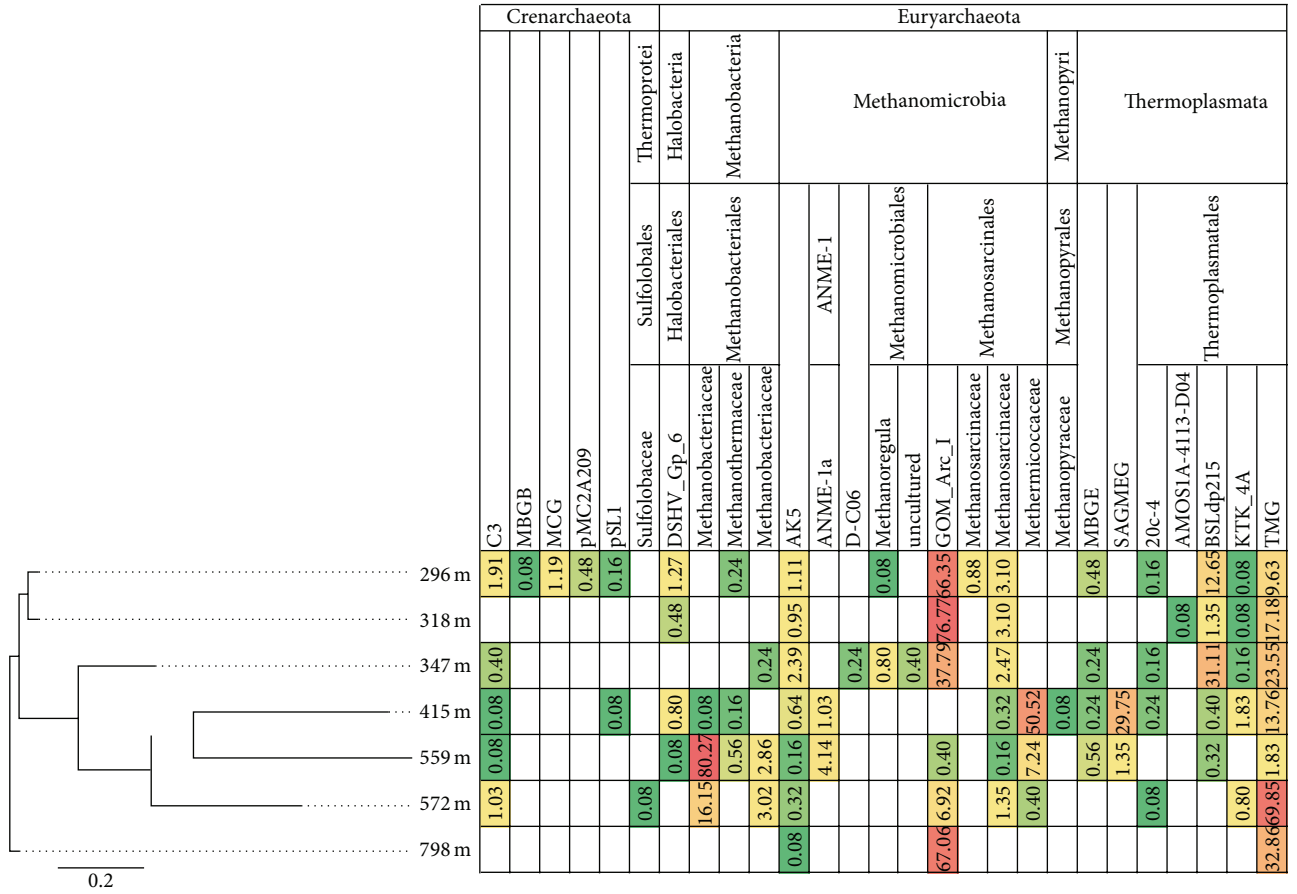


FIGURE 3: The relative distribution of archaeal 16S rRNA sequence reads belonging to specific archaeal families. The relative abundances of sequence reads are highlighted as described in Figure 2. The samples were clustered as described in Figure 2. The data was normalized between the different samples to include 1200 random sequence reads from each sample.

for methane oxidation is the AOM process performed by archaeal ANME linages. Nyyssönen et al. [3] reported putative ANME-1 *mcr*A genes from the 300 to 400 mbgsl in Olkiluoto. In the present study, the active archaeal communities detected in the SMMZ mainly consisted of GOM_Arc_I Methanosarcinales (Figure 3), which also are known as the ANME-2D. ANME-2D archaea have been shown to independently perform nitrate mediated AOM without the need for a bacterial partner [48]. This is in agreement with the *mcr*A gene transcripts detected at this depth, which mostly (55–100%) belonged to Methanosarcinales groups.

At the lower boundaries of the SMMZ at 347 mbgs, the active SRB community changed and the *dsr*B gene transcript pool was dominated by transcripts belonging to an uncultured Desulfobacteraceae group of SRB most closely related to *Desulfobacter* (86.5%), overlapping the distribution of ANME-1 in Olkiluoto. *Desulfosarcina dsr*B transcripts were found only at low abundance but were most numerous at 296–328 mbgsl. Together with the *Desulfobacter* the *Desulfosarcina* also belongs to the Desulfobacteraceae. These *Desulfosarcina* have been reported to form AOM consortia with ANME-1 and ANME-2 archaea [49], which may indicate that these associations also occur in Olkiluoto groundwater SMMZ.

3.2. Methane-Rich Groundwater. Below the SMMZ, at 415–572 mbgsl, the sulphate concentration in the groundwater is greatly reduced, the groundwater salinity increased, and the methane concentration is high. At this depth, γ-proteobacteria most similar to *Pseudomonas* species dominated (41–94%) the active bacterial communities. These bacteria may be the major CO_2-fixing bacteria in Olkiluoto deep methane rich groundwater, as they have been shown to be in the Baltic Sea [50].

A peak in the bacterial diversity was seen at 559 mbgsl in the methane-rich groundwater. Several SMTZ signature groups were detected at this depth including putatively methylotrophic α- and γ-proteobacteria, β-proteobacteria, δ-proteobacterial SRB, JS1, Actinomycetes, Planctomycetes, and Chloroflexi. β-proteobacteria belonging to the Burkholderiales, for example, are believed to be the sole bacterial partner performing nitrification in the AOM association with ANME-2c archaea [51]. β-proteobacterial families Sphingomonadaceae and Comamonadaceae were detected as minority (<3.5%) at all depths. β-proteobacteria were a major group only at 559 mbgsl where *Acidovorax* sp. (Comamonadaceae) contributed almost 15% of the active community and correlated positively and significantly with the highest pH measured in the present study. A low abundance (<1%)

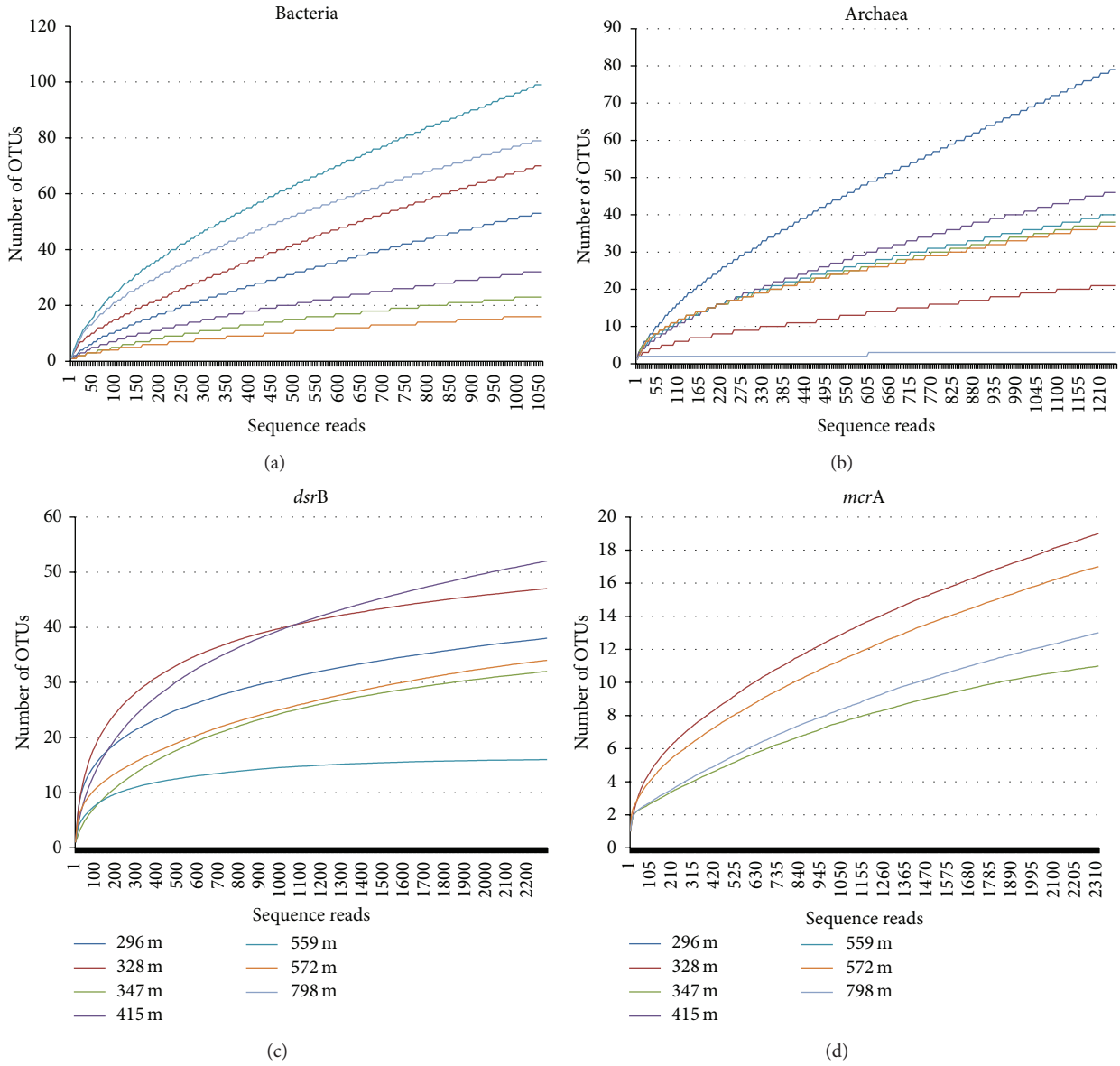


Figure 4: Rarefaction curves of the sequence data obtained from each RNA extract normalized to equal number of sequence reads per sample: (a) bacterial 16S rRNA, (b) archaeal 16S rRNA, (c) *dsr*B transcripts, and (d) *mcr*A transcripts. The x-axis displays the number of sequence reads and the y-axis displays the number of different OTUs obtained. Figures (a)–(c) present rarefaction values at the distance 0.03 and (d) rarefaction for distance 0.01.

of methanotrophic α-proteobacterial Methylobacteriaceae (*Methylobacter* sp. and *Methylocystis* sp.) correlating significantly with depth and salinity were found at this depth. Similar methanotrophs have readily been isolated from anaerobic methane-rich deep subsurface environments, such as terrestrial mud volcanoes [52]. Wrede et al. [52] suggested that aerobic methane oxidation could be activated whenever oxygen was available and thereby keep the subsurface ecosystem anaerobic.

Methylotrophic Methermicoccaceae and SAGMEG Thermoplasmata were the most abundant archaea at 415 mbgsl (50.5% and 29.8%, resp.). Hydrogenotrophic Methanobacteriaceae, which correlated with the highest pH, were the most abundant archaea at 559 mbgsl (80.3%) and terrestrial miscellaneous group (TMG) Thermoplasmatales at 572 mbgsl (69.9%). Nevertheless, the *mcr*A transcripts at 572 mbgsl mostly (75%) belonged to Methanobacteriales methanogens. ANME-1 archaea were found in the methane-rich groundwater at 415 and 559 mbgsl (4.1% and 1.0%, resp.) and correlated positively although not significantly with the highest pH values measured in this study. ANME-1 archaea were most abundant at depths where the GoM_Arch_I/ANME-2D archaea were mainly absent. Recent research shows that some ANME groups are capable of performing sulphate mediated AOM on their own [53], where they form S_2 by a so far unknown sulphate reduction process. The SO_4^{2-}-mediated

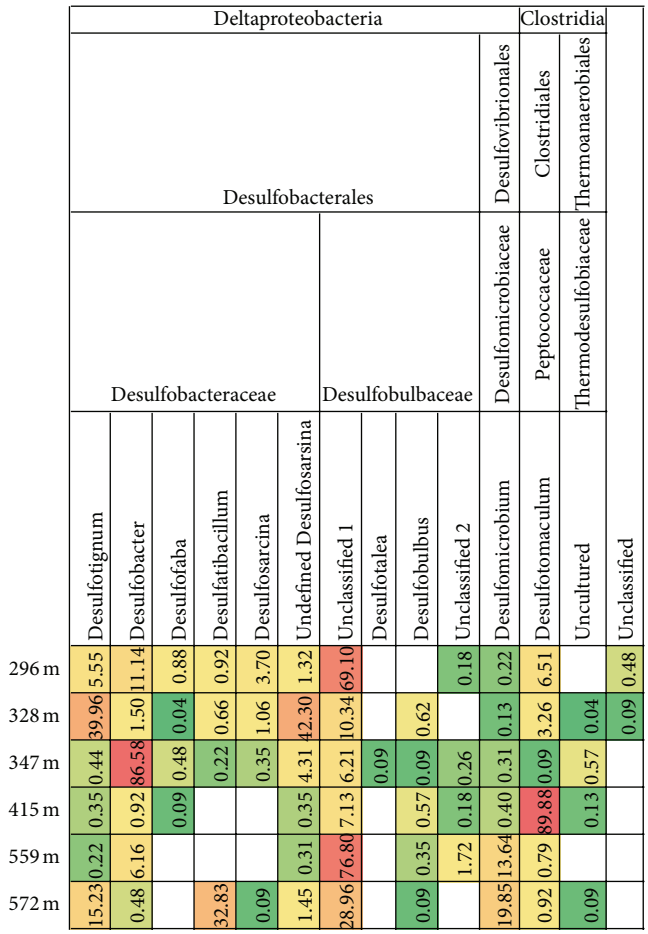


FIGURE 5: The relative distribution of *dsr*B transcript sequence reads belonging to specific SRB families according to the phylogenetic identification of the sequences presented in Figure 6. The relative abundance of sequence reads and the clustering of the samples are presented as described in Figure 2. The data were normalized between the different samples to include 2249 random sequence reads from each sample.

AOM performed by the ANME-1 could dominate specifically at 415–559 mbgsl, where the concentration of methane increases dramatically. Our results are similar to those of Pedersen [54], who suggested that a sulphate mediated AOM process coupled to sulphate reduction may occur in Olkiluoto groundwater at the SMMZ depth, although they did not obtain conclusive evidence for this process.

In the methane-rich water, the methanogens and SRB were clearly enriched at different depths. At 559 mbgsl where the highest number of *mcr*A genes (4.6×10^2 copies mL^{-1}) was detected the number of *dsr*B genes was only 6.5×10^1 copies mL^{-1} and no *dsr*B transcripts could be detected by qPCR. *Dsr*B genes in contrast were abundant above (1.6×10^4 copies mL^{-1} at 415 mbgsl) and below (2.2×10^3 *dsr*B copies mL^{-1} at 572 mbgsl) this depth although the sulphate concentration in the water was only 0.5–1.4 mg L^{-1}. The reason for the higher amount of *dsr*B gene copies mL^{-1} in the sulphate poor water may be that the SRB live by fermentation instead of sulphate reduction. For example, *Desulfobulbus* and *Desulfotomaculum* species have been shown to reduce Fe(III) during fermentation of pyruvate [55, 56]. Both of these sulphate reducers were abundant in the methane rich and sulphate poor groundwater. At 415 mbgsl *Desulfotomaculum dsr*B gene transcripts were the most abundant (89.9%) while Desulfobulbaceae family 1 of the Desulfobacterales dominated (76%) at 559 mbgsl and showed positive and significant correlation with pH the highest groundwater pH. The most even distribution of *dsr*B gene transcripts was seen at 572 mbgsl, where *Desulfatibacillum* (>32%), *Desulfomicrobium* (>19%), and uncultured Desulfobulbaceae (uncultured 1) (>28%) dominated the SRB communities. Firmicutes *dsr*B gene transcripts other than those belonging to *Desulfotomaculum* were detected only at <1% relative abundance at 328 mbgsl and were present at 296–415 mbgsl and 572 mbgsl (Figure 6). These *dsr*B sequences all belonged to *Thermodesulfovibrio* species previously found in soil environments.

3.3. Deep Methane-Rich Groundwater. At 798 mbgsl, the groundwater is highly saline with over 53 g dissolved solids L^{-1} and a high concentration of methane. The microbial community at this depth was clearly different from those at the other depths. However, the bacterial diversity at this depth

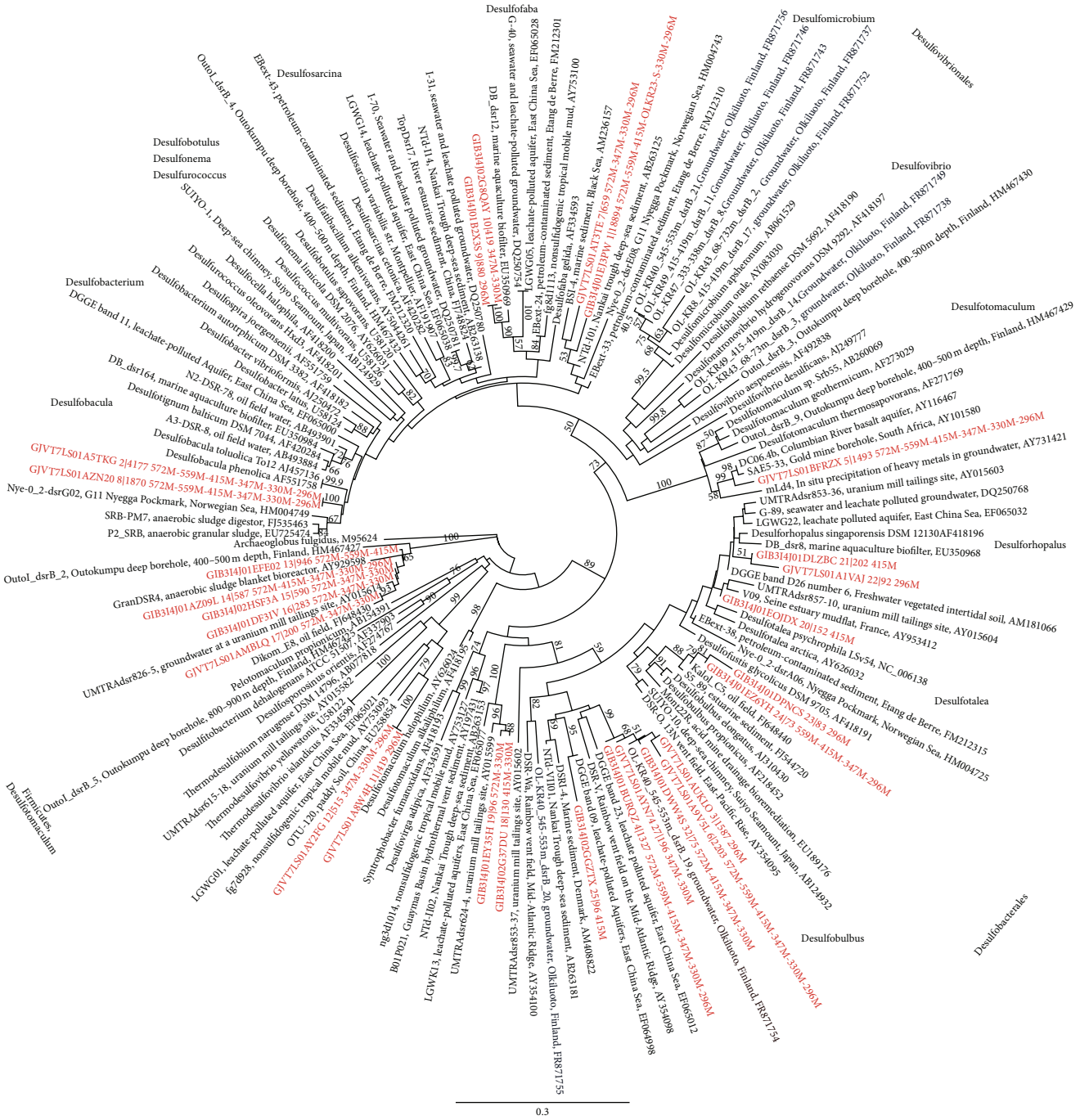


FIGURE 6: The phylogenetic distribution of the amino acid sequences of the OTUs of *dsr*B transcripts detected in this study presented as a maximum likelihood tree. The sequences detected in this study are shown in red. Bootstrap support for nodes was calculated with 1000 random repeats and nodes with more than 50% support are indicated. Sequences detected in this study are shown in red. The sequence name codes consist of the sequence IDENTIFIER and the OTU number|the number of sequence reads in that OTU followed by the depths from which this OTU has been detected.

was surprisingly high. The most common bacterial groups were the Bacillales and Actinobacteria, which significantly increased with increasing depth and salinity throughout the studied depth profile and formed 46% and 18% of the active bacterial community. Archaeal diversity was low, and the archaeal community consisted mainly of GOM_Arc_I Methanosarcinales/ANME-2D (67%) and TMG archaea (32.9%) (Figure 3). No *dsr*B or *mcr*A genes or transcripts were detected by qPCR despite the relatively high microbial density, 2.3×10^4 cells mL^{-1}. In accordance, no *dsr*B

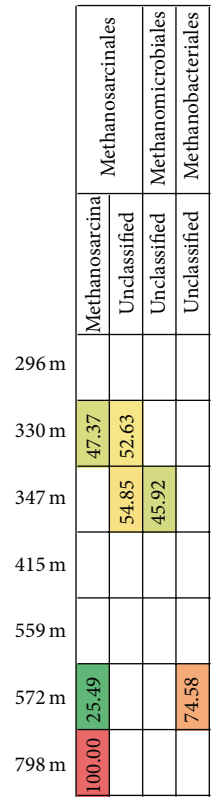


Figure 7: The relative distribution of *mcr*A transcript sequence reads belonging to specific methanogenic archaeal families based on the phylogenetic identification of the *mcr*A reads as presented in Figure 8. The relative abundances of sequence reads are highlighted as described in Figure 2. The data was normalized between the different samples to include 2324 random sequence reads from each sample.

transcripts were obtained for 454 sequencing either. *mcr*A transcripts were obtained by the nested PCR approach only, and all sequences belonged to Methanosarcinales methanogens. The coappearance of these *mcr*A transcripts together with the high relative abundance of GOM_Arc_I Methanosarcinales/ANME-2D archaea indicates active methane cycling activity of GOM_Arc_I Methanosarcinales/ANME-2D archaea at this depth.

4. Conclusions

We observed a clear change in the active microbial community composition at the sulphate-methane interface and the methane-rich groundwater in Olkiluoto. Several SMTZ signature groups were detected, as well as a high diversity of active microorganisms. We found a characteristic increase in the transcription of the *dsr*B gene in the sulphate reducing and putative AOM zone between 296 and 347 mbgsl, coinciding with *mcr*A transcripts of methylotrophic methanogens that possibly belong to the ANME-2D. In methane-rich water between 415 m and 559 mbgsl the ANME-2D were few or absent, while ANME-1 archaea appeared. *mcr*A transcripts from an uncultured group of Methanosarcinales archaea cooccurred with the ANME-2D archaea, but whether they produce or oxidize methane using the reverse methanogenesis pathway is not known.

Overall the active microbial communities in Olkiluoto deep groundwater are diverse and SRB and methanogens are not the only microbial groups to have an influence on hydrogeochemical conditions and to further be taken into account in the safety case of the disposal of spent nuclear fuel. AOM may also be mediated by means other than sulphate or nitrate reduction by different bacterial groups. The great abundance of bacterial and archaeal taxa generally not involved in methane production or oxidation, or nitrate or sulphate reduction, also indicate that the main energy converting metabolic pathways may, in the absence of oxygen, be fermentation of organic molecules.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The research project was funded by Posiva Oy, the Academy of Finland and the Finnish research programme on nuclear waste management (KYT). Mirva Pyrhönen is acknowledged for skillful assistance in the laboratory. Dr. Michael Hardman is thanked for critical language editing.

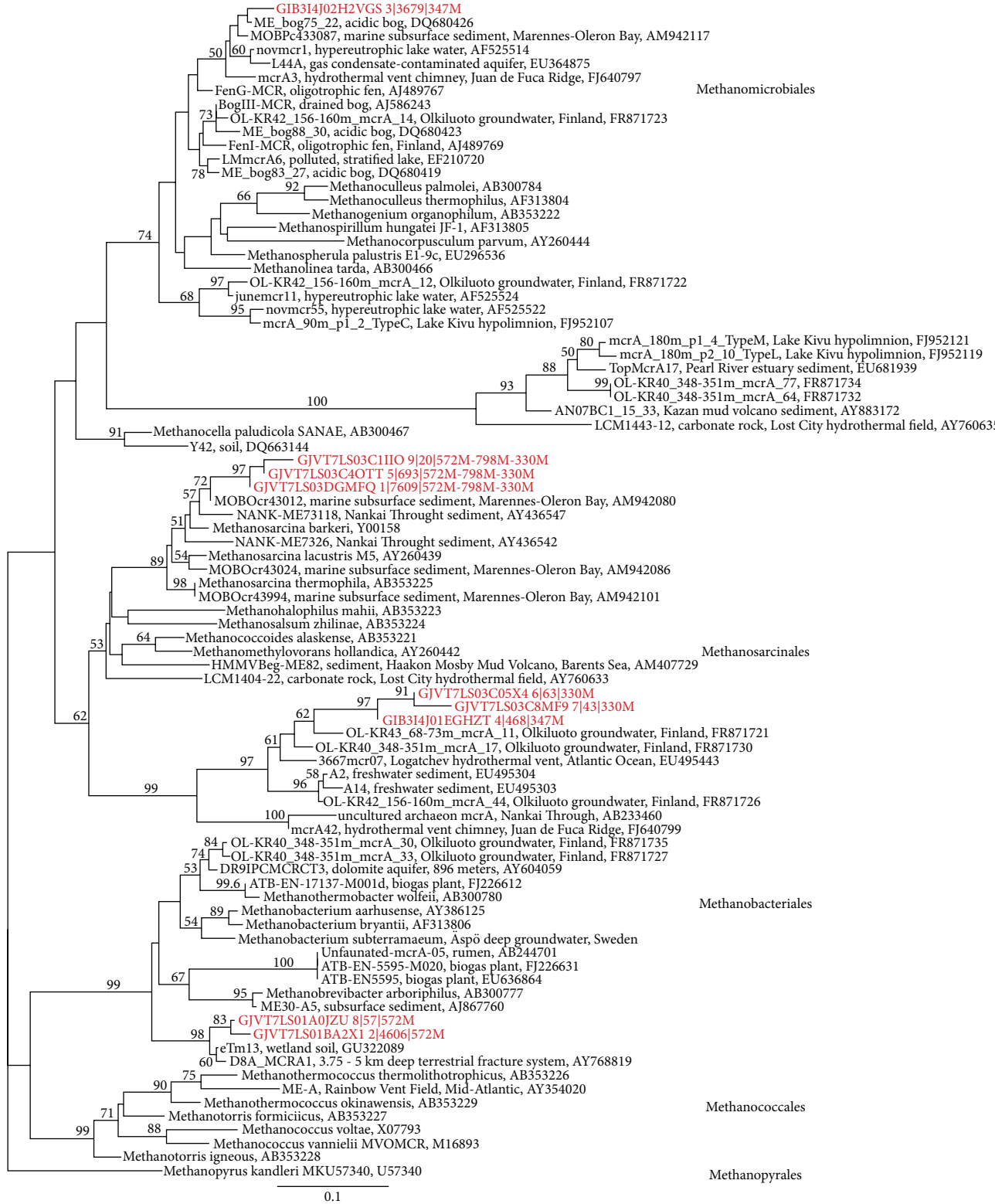


FIGURE 8: The phylogenetic distribution of the amino acid sequences of the OTUs of *mcr*A transcripts obtained detected in this study presented as a maximum likelihood tree. The sequences detected in this study are shown in red. Bootstrap support for nodes was calculated with 1000 random repeats and nodes with more than 50% support are indicated. The sequence codes are as described in Figure 6.

Table 4: The number of sequence reads, the observed and estimated number of OTUs, diversity coverage, and diversity index (H') obtained by the HTP sequencing of bacterial and archaeal 16S rRNA and *dsr*B and *mcr*A transcripts. The diversity and OTU richness estimates were calculated based on equal number of sequence reads.

		296 m	328 m	347 m	415 m	559 m	572 m	798 m
Bacteria 16S	Number of reads	1220	893	9209	18425	17158	5377	996
	Observed OTUs	45	46	35	63	83	169	49
	Estimated richness							
	Chao	161	294	73	126	161	367	104
	Ace	412	355	111	219	355	587	139
	Coverage % chao	28	16	48	50	56	46	47
	$^{*}H'$	1.67	1.25	0.15	0.44	1.79	0.17	1.43
Archaea 16S	Number of reads	6322	12785	1377	2122	1223	1655	3277
	Observed OTUs	139	67	32	45	26	26	6
	Estimated richness							
	Chao	249	137	47	46	46	28	7
	Ace	382	210	81	60	75	30	16
	Coverage % chao	56	49	68	98	57	93	86
	$^{*}H'$	1.06	0.81	1.91	1.23	0.80	1.04	0.83
*dsr*B	Number of reads	8131	4144	12649	8628	2360	2249	—
	Observed OTUs	33	41	47	50	13	26	
	Estimated richness							
	Chao	38	42	51	60	16	31	
	Ace	39	45	52	63	38	49	
	Coverage % chao	86.8	97.6	92.2	83.3	81.3	83.9	
	$^{*}H'$	2.29	2.65	0.69	1.03	1.93	1.81	
*mcr*A	Number of reads	—	4188	4184	—	—	6676	2324
	Observed OTUs		4	2			2	1
	Estimated richness							
	Chao		4	2			2	1
	Ace		5	0			0	0
	Coverage % chao		100	100			100	100
	$^{*}H'$		0.45	0.42			0.76	0.43

*Normalized according to sample with the lowest number of reads.

References

[1] H. Castaneda and X. D. Benetton, "SRB-biofilm influence in active corrosion sites formed at the steel-electrolyte interface when exposed to artificial seawater conditions," *Corrosion Science*, vol. 50, no. 4, pp. 1169–1183, 2008.

[2] Posiva Oy, "Olkiluoto site description 2011," Report POSIVA 2011-02, Posiva Oy, 2012.

[3] M. Nyyssönen, M. Bomberg, A. Kapanen, A. Nousiainen, P. Pitkänen, and M. Itävaara, "Methanogenic and sulphate-reducing microbial communities in deep groundwater of crystalline rock fractures in Olkiluoto, Finland," *Geomicrobiology Journal*, vol. 29, no. 10, pp. 863–878, 2012.

[4] S. A. Haveman, K. Pedersen, and P. Ruotsalainen, "Distribution and metabolic diversity of microorganisms in deep igneous rock aquifers of Finland," *Geomicrobiology Journal*, vol. 16, no. 4, pp. 277–294, 1999.

[5] S. A. Haveman and K. Pedersen, "Distribution of culturable microorganisms in Fennoscandian shield groundwater," *FEMS Microbiology Ecology*, vol. 39, no. 2, pp. 129–137, 2002.

[6] M. Bomberg, M. Nyyssönen, A. Nousiainen et al., "Evaluation of molecular techniques in characterization of deep terrestrial biosphere," *Open Journal of Ecology*, vol. 4, pp. 468–487, 2014.

[7] L. Hallbeck and K. Pedersen, "Culture-dependent comparison of microbial diversity in deep granitic groundwater from two sites considered for a Swedish final repository of spent nuclear fuel," *FEMS Microbiology Ecology*, vol. 81, no. 1, pp. 66–77, 2012.

[8] M. Itävaara, M. Nyyssönen, A. Kapanen, A. Nousiainen, L. Ahonen, and I. Kukkonen, "Characterization of bacterial diversity to a depth of 1500m in the Outokumpu deep borehole, Fennoscandian Shield," *FEMS Microbiology Ecology*, vol. 77, no. 2, pp. 295–309, 2011.

[9] R. J. Parkes, B. A. Cragg, N. Banning et al., "Biogeochemistry and biodiversity of methane cycling in subsurface marine sediments (Skagerrak, Denmark)," *Environmental Microbiology*, vol. 9, no. 5, pp. 1146–1161, 2007.

[10] G. Webster, H. Sass, B. A. Cragg et al., "Enrichment and cultivation of prokaryotes associated with the sulphate-methane transition zone of diffusion-controlled sediments of Aarhus Bay, Denmark, under heterotrophic conditions," *FEMS Microbiology Ecology*, vol. 77, no. 2, pp. 248–263, 2011.

[11] J. Leloup, A. Loy, N. J. Knab, C. Borowski, M. Wagner, and B. B. Jørgensen, "Diversity and abundance of sulfate-reducing

microorganisms in the sulfate and methane zones of a marine sediment, Black Sea," *Environmental Microbiology*, vol. 9, no. 1, pp. 131–142, 2007.

[12] L. Holmkvist, T. G. Ferdelman, and B. B. Jørgensen, "A cryptic sulfur cycle driven by iron in the methane zone of marine sediment (Aarhus Bay, Denmark)," *Geochimica et Cosmochimica Acta*, vol. 75, no. 12, pp. 3581–3599, 2011.

[13] N. Iversen and B. B. Jorgensen, "Anaerobic methane oxidation rates at the sulfate-methane transition in marine sediments from Kattegat and Skagerrak (Denmark)," *Limnology & Oceanography*, vol. 30, no. 5, pp. 944–955, 1985.

[14] Y. Koizumi, S. Takii, M. Nishino, and T. Nakajima, "Vertical distributions of sulfate-reducing bacteria and methane-producing archaea quantified by oligonucleotide probe hybridization in the profundal sediment of a mesotrophic lake," *FEMS Microbiology Ecology*, vol. 44, no. 1, pp. 101–108, 2003.

[15] P. Aalto, J. Helin, S. Lindgren et al., *Baseline Report for Infiltration Experiment*, WR 2011–25, Posiva Oy, Olkiluoto, Finland, 2011.

[16] K. Pedersen, A. F. Bengtsson, J. S. Edlund, and L. C. Eriksson, "Sulphate-controlled diversity of subterranean microbial communities over depth in deep groundwater with opposing gradients of sulphate and methane," *Geomicrobiology Journal*, vol. 31, no. 7, pp. 617–631, 2014.

[17] K. Pedersen, J. Arlinger, S. Erikson et al., *Microbiology of Olkiluoto Groundwater, Results and Interpretations 2007*, WR 2008-34, 2008.

[18] Posiva, "Olkiluoto site description 2008, part 1," POSIVA 2009-01, Posiva Oy, 2009.

[19] A. Kärki and S. Paulamäki, "Petrology of Olkiluoto," POSIVA 2006-02, Posiva Oy, 2006.

[20] H. Ahokas, E. Tammisto, and T. Lehtimäki, "Baseline head in Olkiluoto," WR 2008-69, Posiva Oy, Eurajoki, Finland, 2008.

[21] R. L. Kepner and J. R. Pratt, "Use of fluorochromes for direct enumeration of total bacteria in environmental samples: past and present," *Microbiological Reviews*, vol. 58, no. 4, pp. 603–615, 1994.

[22] M. Itävaara, M.-L. Vehkomäki, and A. Nousiainen, "Sulphate-reducing bacteria in ground water samples from Olkiluoto—analyzed by quantitative PCR," Tech. Rep. WR 2008-82, 2008.

[23] U. Edwars, T. Rogall, H. Blocker, M. Emde, and E. C. Bottger, "Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA," *Nucleic Acids Research*, vol. 17, no. 19, pp. 7843–7853, 1989.

[24] G. Muyzer, E. C. de Waal, and A. G. Uitterlinden, "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA," *Applied and Environmental Microbiology*, vol. 59, no. 3, pp. 695–700, 1993.

[25] R. Großkopf, P. H. Janssen, and W. Liesack, "Diversity and structure of the methanogenic community in anoxic rice paddy soil microcosms as examined by cultivation and direct 16S rRNA gene sequence retrieval," *Applied and Environmental Microbiology*, vol. 64, no. 3, pp. 960–969, 1998.

[26] D. A. Stahl and R. Amann, "Development and application of nucleic acid probes in bacterial systematics," in *Nucleic Acid Techniques in Bacterial Systematics*, E. Stackebrandt and M. Goodfellow, Eds., pp. 205–248, John Wiley and Sons, New York, NY, USA, 1998.

[27] N. Bano, S. Ruffin, B. Ransom, and J. T. Hollibaugh, "Phylogenetic composition of Arctic ocean archaeal assemblages and comparison with Antarctic assemblages," *Applied and Environmental Microbiology*, vol. 70, no. 2, pp. 781–789, 2004.

[28] S. M. Barns, R. E. Fundyga, M. W. Jeffries, and N. R. Pace, "Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 5, pp. 1609–1613, 1994.

[29] J. Geets, B. Borremans, L. Diels et al., "DsrB gene-based DGGE for community and diversity surveys of sulfate-reducing bacteria," *Journal of Microbiological Methods*, vol. 66, no. 2, pp. 194–205, 2006.

[30] M. Wagner, A. J. Roger, J. L. Flax, G. A. Brusseau, and D. A. Stahl, "Phylogeny of dissimilatory sulfite reductases supports an early origin of sulfate respiration," *Journal of Bacteriology*, vol. 180, no. 11, pp. 2975–2982, 1998.

[31] B. A. Hales, C. Edwards, D. A. Ritchie, G. Hall, R. W. Pickup, and J. R. Saunders, "Isolation and identification of methanogen-specific DNA from blanket bog peat by PCR amplification and sequence analysis," *Applied and Environmental Microbiology*, vol. 62, no. 2, pp. 668–675, 1996.

[32] P. D. Schloss, S. L. Westcott, T. Ryabin et al., "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[33] E. Pruesse, C. Quast, K. Knittel et al., "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Research*, vol. 35, no. 21, pp. 7188–7196, 2007.

[34] V. Zverlov, M. Klein, S. Lücker et al., "Lateral gene transfer of dissimilatory (bi)sulfite reductase revisited," *Journal of Bacteriology*, vol. 187, no. 6, pp. 2203–2208, 2005.

[35] T. Z. DeSantis, P. Hugenholtz, N. Larsen et al., "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," *Applied and Environmental Microbiology*, vol. 72, no. 7, pp. 5069–5072, 2006.

[36] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[37] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[38] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," in *Mammalian Protein Metabolism*, H. N. Munro, Ed., pp. 21–132, Academic Press, New York, NY, USA, 1969.

[39] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach," *Molecular Biology and Evolution*, vol. 18, no. 5, pp. 691–699, 2001.

[40] Ø. Hammer, D. A. T. Harper, and P. D. Ryan, "Past: paleontological statistics software package for education and data analysis," *Palaeontologia Electronica*, vol. 4, no. 1, p. 9, 2001.

[41] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 1965.

[42] M. A. Stephens, "Tests based on EDF statistics," in *Goodness-of-Fit Techniques*, R. B. D'Agostino and M. A. Stephens, Eds., pp. 97–194, Marcel Dekker, New York, NY, USA, 1986.

[43] S. D'Hondt, S. Rutherford, and A. J. Spivack, "Metabolic activity of subsurface life in deep-sea sediments," *Science*, vol. 295, no. 5562, pp. 2067–2070, 2002.

[44] I. Roalkvam, S. L. Jørgensen, Y. Chen et al., "New insight into stratification of anaerobic methanotrophs in cold seep sediments," *FEMS Microbiology Ecology*, vol. 78, no. 2, pp. 233–243, 2011.

[45] J. Grote, *Physiology, ecology, and genomics of facultative chemoautotrophic Epsilonproteobacteria in marine pelagic redoxclines [Ph.D. thesis]*, University of Rostock, Rostock, Germany, 2009.

[46] K. Takai, M. Suzuki, S. Nakagawa et al., "*Sulfurimonas paralvinellae* sp. nov., a novel mesophilic, hydrogen- and sulfur-oxidizing chemolithoautotroph within the *Epsilonproteobacteria* isolated from a deep-sea hydrothermal vent polychaete nest, reclassification of *Thiomicrospira denitrificans* as *Sulfurimonas denitrificans* comb. nov. and emended description of the genus *Sulfurimonas*," *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, no. 8, pp. 1725–1733, 2006.

[47] S. Glaubitz, T. Lueders, W.-R. Abraham, G. Jost, K. Jürgens, and M. Labrenz, "^{13}C-isotope analyses reveal that chemolithoautotrophic Gamma- and Epsilon-proteobacteria feed a microbial food web in a pelagic redoxcline of the central Baltic Sea," *Environmental Microbiology*, vol. 11, no. 2, pp. 326–337, 2009.

[48] M. F. Haroon, S. Hu, Y. Shi et al., "Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage," *Nature*, vol. 500, no. 7464, pp. 567–570, 2013.

[49] V. J. Orphan, C. H. House, K.-U. Hinrichs, K. D. McKeegan, and E. F. DeLong, "Multiple archaeal groups mediate methane oxidation in anoxic cold seep sediments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 11, pp. 7663–7668, 2002.

[50] S. Glaubitz, T. Lueders, W. R. Abraham, G. Jost, K. Jürgens, and M. Labrenz, "13C-isotope analyses reveal that chemolithoautotrophic Gamma- and *Epsilonproteobacteria* feed a microbial food web in a pelagic redoxcline of the central Baltic Sea," *Environmental Microbiology*, vol. 11, no. 2, pp. 326–337, 2009.

[51] K. Knittel and A. Boetius, "Anaerobic oxidation of methane: progress with an unknown process," *Annual Review of Microbiology*, vol. 63, pp. 311–334, 2009.

[52] C. Wrede, A. Dreier, S. Kokoschka, and M. Hoppert, "Archaea in symbioses," *Archaea*, vol. 2012, Article ID 596846, 11 pages, 2012.

[53] J. Milucka, T. G. Ferdelman, L. Polerecky et al., "Zero-valent sulphur is a key intermediate in marine methane oxidation," *Nature*, vol. 491, no. 7425, pp. 541–546, 2012.

[54] K. Pedersen, "Metabolic activity of subterranean microbial communities in deep granitic groundwater supplemented with methane and H_2," *The ISME Journal*, vol. 7, no. 4, pp. 839–849, 2013.

[55] D. R. Lovley, E. E. Roden, E. J. P. Phillips, and J. C. Woodward, "Enzymatic iron and uranium reduction by sulfate-reducing bacteria," *Marine Geology*, vol. 113, no. 1-2, pp. 41–53, 1993.

[56] E. Dalla Vecchia, E. I. Suvorova, J. Maillard, and R. Bernier-Latmani, "Fe(III) reduction during pyruvate fermentation by *Desulfotomaculum reducens* strain MI-1," *Geobiology*, vol. 12, no. 1, pp. 48–61, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 892013, 12 pages
http://dx.doi.org/10.1155/2015/892013

Research Article

454-Pyrosequencing Analysis of Bacterial Communities from Autotrophic Nitrogen Removal Bioreactors Utilizing Universal Primers: Effect of Annealing Temperature

Alejandro Gonzalez-Martinez,[1] Alejandro Rodriguez-Sanchez,[2] Belén Rodelas,[2] Ben A. Abbas,[3] Maria Victoria Martinez-Toledo,[2] Mark C. M. van Loosdrecht,[3] F. Osorio,[1] and Jesus Gonzalez-Lopez[2]

[1]*Department of Civil Engineering, University of Granada, Campus de Fuentenueva, s/n, 18071 Granada, Spain*
[2]*Institute of Water Research, University of Granada, C/Ramón y Cajal 4, 18071 Granada, Spain*
[3]*Department of Biotechnology, Technical University of Delft, Julianalaan 67, 2628 BC Delft, Netherlands*

Correspondence should be addressed to Alejandro Gonzalez-Martinez; agon@ugr.es

Received 20 October 2014; Accepted 26 February 2015

Academic Editor: Weixing Feng

Identification of anaerobic ammonium oxidizing (anammox) bacteria by molecular tools aimed at the evaluation of bacterial diversity in autotrophic nitrogen removal systems is limited by the difficulty to design universal primers for the *Bacteria* domain able to amplify the anammox 16S rRNA genes. A metagenomic analysis (pyrosequencing) of total bacterial diversity including anammox population in five autotrophic nitrogen removal technologies, two bench-scale models (MBR and Low Temperature CANON) and three full-scale bioreactors (anammox, CANON, and DEMON), was successfully carried out by optimization of primer selection and PCR conditions (annealing temperature). The universal primer 530F was identified as the best candidate for total bacteria and anammox bacteria diversity coverage. Salt-adjusted optimum annealing temperature of primer 530F was calculated (47°C) and hence a range of annealing temperatures of 44–49°C was tested. Pyrosequencing data showed that annealing temperature of 45°C yielded the best results in terms of species richness and diversity for all bioreactors analyzed.

1. Introduction

Anaerobic ammonium oxidizing (anammox) bacteria belong to the *Candidatus* Brocadiales order first described in 1999 [1]. Their ability to perform anaerobic ammonium oxidation has attracted the attention of many researchers due to the change it made for the understanding of the nitrogen cycle. They have been found as important bacteria for the ecology of nitrogen in oceanic environments [2–4] and have been proposed to play an important role in the nitrogen cycle at global scale [2]. Anammox bacteria have been identified in many natural and engineered environments such as marine sediments, agricultural soils, or wastewater treatment plants [5–7]. In addition, anammox bacteria are the basis for promising technologies aimed at removing nitrogen from wastewater. The autotrophic, anaerobic ammonium oxidation has been utilized for the set-up of several technologies for nitrogen removal, such as partial nitritation/anammox, DEMON, OLAND, or CANON [8]. Autotrophic nitrogen removal technologies account for several advantages over traditional total nitrification-denitrification processes, such as the lesser bioreactor volume required, lower biomass production, or saving costs in aeration and carbon source requirements [9–11]. Thus, anammox bacteria have become of relevance in both natural and engineered systems, and more research about the ecology of the ecosystems where they develop is expected in the following years.

PCR-based molecular biology techniques such as qPCR quantification and high-throughput pyrosequencing are powerful tools for the estimation of the abundance

and diversity of microorganisms in natural and engineered ecosystems [12, 13]. Most frequently, anammox-specific primers have been applied for PCR-based evaluation of the occurrence and diversity of these organisms [12, 14–18]. The reason under this common practice is that universal primers targeting the whole *Bacteria* domain tend to amplify poorly the 16S rRNA gene of anammox due to their not very high identity (<87.1%) [19, 20]. In this sense, when universal primers are used PCR amplification of some anammox phylotypes might be missing, with the consequent loss of fundamental information on the microbial diversity of the system. Nevertheless, the use of anammox-specific primers is not sufficient for a complete understanding of microbial ecosystems, given that other bacteria are not taken into account.

In this research, several universal primers and annealing temperature conditions for PCR amplification were tested in order to achieve the best combination possible for its use on the metagenomic analysis of ecosystems, such as autotrophic nitrogen removal bioreactors. After an *in silico* testing, a universal primer was selected for the best coverage of 16S rRNA genes of the domain *Bacteria* including anammox (*Planctomycetes*). Using this primer, bacterial diversity in samples from five different lab-scale and full-scale autotrophic nitrogen removal bioreactors (see Table S1 in Supplementary Materials available online at http://dx.doi.org/10.1155/2015/892013) were analyzed by high-throughput pyrosequencing, using six different annealing temperatures. Based on the results obtained, a robust method for the evaluation of the bacterial diversity in ecosystems where anammox bacteria are of importance is proposed.

2. Materials and Methods

2.1. Bioreactors. Five different autotrophic nitrogen removal bioreactors which represent the main existing anammox technologies were sampled in the analysis (Table S1). Two of them were bench-scale models and the other three were full-scale plant bioreactors. The bench-scale models analyzed in the study were named Lab MBR and Low Temperature CANON. Lab MBR is a membrane bioreactor (MBR) anammox system and Low Temperature CANON is a CANON system operated at 15°C. Both bioreactors were built in Netherlands in 2010 and 2009, respectively, and were fed synthetic wastewater.

The full-scale bioreactors sampled in the study were located at three cities in Netherlands, Apeldoorn, Olburgen, and Rotterdam, and were named A, N, and R, respectively. Bioreactor A is a DEMON system built in 2010 which treats reject water from anaerobic digester. Bioreactor N is a CANON process treating sewage from a potato processing factory built in 2009. Bioreactor R is the anammox reactor in a two-step anammox plant constructed in 2002 and treats reject water from anaerobic digester.

2.2. Sampling. Five samples (200 mL) were taken from five evenly distributed points within each bioreactor volume. The procedures for sampling and pretreatment for DNA extraction followed those described by Ni et al. [16]. Biomass was separated from the collected wastewater by centrifugation at 3500 rpm for 10 minutes at room temperature in a Kokusan H-103N series apparatus (Kokusan Enshinki Co., Ltd., Tokyo, Japan). Pelleted biomass was stored at −20°C before DNA extraction.

2.3. In Silico Primer Evaluation. The *primer pair* 530F (5′-GTGCCAGCMGCNGCGG)-1100R (5′-GGGTTNCGNTCGTTG), described by Dowd et al. [21], was selected after an *in silico* analysis of several universal primers targeting the 16S rRNA gene and commonly used in earlier literature. The primers were tested by correlating the accession numbers of the matching sequences, using the Probe Match function of the Ribosomal Data Project (RDP) (http://rdp.cme.msu.edu/probematch/search.jsp) (see Supplementary Material). Moreover, to double check the performance of the selected primer set over the *Planctomycetes* phylum, an extra *in silico* analysis study using the SiLVA Test Probe tool was done (http://www.arb-silva.de/search/testprobe).

2.4. In Silico Calculation of Optimum Annealing Temperature. Calculation of optimum annealing temperature for primer 530F was done following the expression for the optimum salt-adjusted annealing temperature of a primer developed by Rychlik et al. [22–25]. For calculation purposes, combined concentration of Na^+ and K^+ was taken as 50 mM.

2.5. DNA Extraction, PCR Amplification, and Pyrosequencing. 300 mg of pelleted biomass for each centrifuged sample was used for DNA extraction using the Fast DNA SPIN Kit for Soil and the Fast-Prep24 apparatus (MP Biomedicals, Solon, OH, USA) following the instructions given by the manufacturer. DNA extracts from samples collected in the same bioreactor were merged into a pool. The DNA pool of each bioreactor was divided into 6 subsamples with equal volume for further pyrosequencing analysis.

The primer pair 530F-1100R [21] was used to amplify 500 bp of the 16S rRNA gene of *Bacteria*, encompassing the V4-V5-V6 hypervariable regions. Research and Testing Laboratory (Lubbock, Texas, USA) proceeded with pyrosequencing following the procedure described by Dowd et al., 2008 [21], using the Roche 454 GS-FLX+ apparatus. Amplification of the six subsamples within the same bioreactor DNA pool was assayed under the same PCR conditions but at different annealing temperatures (44 to 49°C), yielding a total of 30 different pyrosequencing datasets in the five different technologies. In this sense the PCR conditions for pyrosequencing were the following: preheating step at 94°C for 3 minutes; 32 cycles at 94°C for 30 seconds, 44–49°C for 40 seconds, and 72°C for 1 minute; elongation at 72°C for 5 minutes.

2.6. Pyrosequencing Postrun Analysis. Elimination of poor-quality end reads from pyrosequencing raw data was done by quality trimming based on quality scores. USEARCH [26] was then used to generate seed sequences to which

quality trimmed reads were clustered within a 4% divergence threshold. This procedure eliminates sequences that fail to encounter similar enough reads. Chimera detection was developed using the *de novo* method implemented in UCHIIME [27] over clustered quality trimmed sequences collected during the previous step. Denoising was then conducted for elimination of bad sequences and correction of base pair errors. Following denoising, a quality control screening was conducted. Quality criteria taken were the following: (1) failed sequence reads, (2) sequences with low quality tags, and (3) sequences that are shorter than half the expected amplicon length or 250 bp, whichever the shortest. Sequences that could not meet the defined quality criteria were eliminated. Reads that passed the quality screening control were then clustered for phylogenetic identification into 0% divergence using USEARCH [26]. The Kraken BLAST software (http://ccb.jhu.edu/software/kraken/) [28] was utilized to provide a seed sequence for each phylogenetic identification cluster from a high-quality database derived from the NCBI GenBank database. Based on BLASTN+ identity, sequences were affiliated to distinct taxonomic levels as follows: (1) at OTU level if divergence was less than 3%, (2) at genus level if divergence was in the range 3–5%, (3) at family level if divergence was 5–10%, (4) at order level if divergence was 10–15%, (5) at class level if divergence was 15–20%, and (6) at phylum level if divergence was 20–23%. Sequences that did not encounter queried sequences with less than 23% divergence were discarded.

2.7. Rarefaction Curves. Rarefaction curves for each sample were calculated using the aRarefactWin software developed by S. Holland (University of Georgia, Athens; http://strata.uga.edu/software/). For the purpose of microbial community analysis, rarefaction curves of samples belonging to the same bioreactor were interpolated to the lowest reads count among all of them and extrapolated to the highest reads count among all of them. Extrapolation of rarefaction curves was done following the mathematical model described in Colwell et al. [29].

2.8. Heat Maps. For each bioreactor, a heat map was generated defining the differences of community structure on the basis of the annealing temperatures selected for PCR. The heat maps were based on the relative abundance of OTUs > 1% relative abundance.

2.9. Cluster Analysis (CA). Three different types of cluster analysis (CA) were performed over the samples: "analysis A" including all OTUs and using a non-phylogeny-dependent method; "analysis B" including only OTUs with >1% relative abundance and using a non-phylogeny-dependent method; and "analysis C" including only OTUs with >1% relative abundance and using a phylogeny-dependent method. Non-phylogeny-dependent A and B analysis were based on Bray-Curtis dissimilarity and carried out with the Vegan package v.2.0 implemented on the statistical software R-Project v.2.15.1 [30]. The software Fast UniFrac [31] was utilized for the phylogeny-dependent C analysis, following instructions given by the developers in the software tutorial (http://unifrac.colorado.edu/). For this purpose, reference trees were constructed for each bioreactor, using the MEGA 6.0 software [32]. Relative abundances of each of the OTUs were used as weight for the analysis.

2.10. Principal Coordinates Analysis (PCoA). Principal coordinates analysis (PCoA) of samples coming from the same bioreactor was done based on a phylogeny-dependent method. OTUs with relative abundance >1% were selected for the study. For each bioreactor, a reference tree with selected OTUs was constructed using MEGA 6.0 software. Utilizing reference trees, phylogeny-based PCoA was conducted using Fast UniFrac software [31], following the instructions given in the software tutorial (http://unifrac.colorado.edu/).

2.11. Hill Diversity Indices. Hill diversity indices of order 1 (Shannon index) and order 2 (Simpson index) were calculated for each sample using the Vegan package v.2.0 implemented on the statistical software R-Project v.2.15.1 [30].

3. Results and Discussion

3.1. In Silico Primer Evaluation. The design of a primer that covers all species within the domain *Bacteria* is impossible [33]. Therefore, when a primer is chosen for a bacterial community analysis, one has to accept that no total coverage could be found. For this reason, potential coverage of a primer is of primary importance in order to decide the best option for the analysis of microbial ecology of a natural ecosystem. Results of the *in silico* search for coverage of bacterial species of nine widely used universal primers within the RDP database are displayed in Table 1. Only results for the phylum *Planctomycetes* and the complete *Bacteria* domain are shown. Results for all other phyla can be seen in Table S2. The results of the *in silico* analysis showed that primer 530F offered the best coverage of species for both the *Bacteria* domain and the phylum *Planctomycetes*. Primer 530F was able to discern 76.01% of the total bacterial species known to date. The closest follower was primer 519R, with 8% less total coverage. With regard to *Planctomycetes*, primer 530F covered 75.58% of species belonging to this group, while the second best (910R) covered 53.70%.

Coverage of all phylotypes of anammox bacteria has been reported as very difficult if a universal primer is utilized [19, 20]. *In silico* analysis of the coverage of all *Candidatus* Brocadiales microorganisms by primer 530F within the SiLVA database is summarized in Table S3. We found that primer 530F targeted the 16S rRNA genes of all the anammox *Candidatus* species known to date yielding a minimum of 90% perfect matches to sequences filed in the database. Thus, it can be stated that primer 530F is the best option for the metagenomic analysis of autotrophic nitrogen removal bioreactors.

For the purposes of pyrosequencing, reverse primer 1100R was taken, following the literature that has utilized primer 530F for pyrosequencing analysis [21].

Table 1: *In silico* evaluation of primer coverage of species in the *Bacteria* domain and *Planctomycetes* specific.

Phylum	27F	519R	530F	787R	910R	1064R	1392R	1492R
Planctomycetes	17.82%	23.89%	75.58%	6.17%	53.70%	39.88%	33.25%	6.59%
Total	**11.58%**	**68.47%**	**76.01%**	**59.45%**	**56.53%**	**54.52%**	**22.08%**	**4.08%**

3.2. In Silico Optimum Annealing Temperature Calculation. Following the expression given by Rychlik et al. [22] for the salt-adjusted optimum melting temperature of primer 530F with combined N^+ and K^+ concentration of 50 mM, as utilized by other authors [23–25], optimum melting temperature for primer 530F was 47°C. For the testing of different annealing temperatures an interval around optimum annealing temperature calculated was covered, ranging from 44°C to 49°C.

3.3. Rarefaction Curves. For each bioreactor, original rarefaction curves, interpolated ones to lowest number of reads within samples, and extrapolated ones to the highest number of reads within samples were generated and are shown in Figure S1. It can be seen that annealing temperatures of 44°C offer lower species richness than all the other annealing temperatures tested.

3.4. Diversity and Relative Abundance of Bacterial Species. Heat maps were generated for each bioreactor taking into account only the OTUs with >0.1% relative abundance as shown in Figures 1 and 2. Annealing temperature of 44°C showed a pattern that differed from those generated at all other temperatures in all bioreactors. Interestingly, at 44°C bacteria belonging to the *Candidatus* Brocadiales order (*Candidatus* Brocadia anammoxidans, *Brocadia fulgida,* and *Brocadia* sp.) showed a much lower relative abundance than that at all the other annealing temperatures tested for every bioreactor. Particularly, for the Lab MBR bioreactor, which stands as a highly enriched (>90%) [34] culture of Brocadiales, the relative abundance found was 1.27% when 44°C was the annealing temperature, while at the other temperatures assayed it fell in the 75–86% range.

Amplification of partial 16S rRNA genes of bacteria other than anammox was also influenced by the annealing temperature selected. In the case of the Lab MBR, the relative abundance of copies of *Carboxydibrachium* sp.was of 84.5% at 44°C, much higher compared to the other annealing temperatures tested. The inability to capture any *Candidatus* Brocadiales phylotypes in all bioreactors, with one of those being a high enrichment of these bacteria, when annealing was performed at 44°C is related to a consistent error occurring during the PCR procedure of primer 530F. In this sense, it is possible that annealing at 44°C hinds the relative abundance of *Candidatus* Brocadiales bacteria under other bacterial species such as *Carboxydibrachium* sp. in the different pyrosequencing samples from the five bioreactors analyzed.

The results described here demonstrated that an annealing temperature of 44°C was inappropriate to analyze the bacterial community structure of the bioreactors sampled in the study, due to a severe underestimation of the relative abundance of Brocadiales.

Results of bacterial diversity obtained for bioreactors Lab MBR and Low Temperature CANON were further analyzed to check the performance of primer 530F for the evaluation of anammox diversity. Diversity of anammox bacteria in bioreactors Lab MBR and Low Temperature CANON was studied before by other authors utilizing different molecular biology techniques, such as FISH or qPCR [35, 36]. In these studies, it was concluded that Candidatus *Brocadia fulgida* was the only anammox bacteria in both Lab MBR and Low Temperature CANON bioreactors. In the present study, Candidatus *Brocadia* sp. was the only anammox identified in Lab MBR bioreactor, while in the Low Temperature CANON bioreactor, pyrosequencing with primer 530F retrieved partial 16S rRNA genes affiliated to four different anammox phylotypes: *Candidatus* Brocadia anammoxidans, *Candidatus Brocadia fulgida*, *Candidatus Brocadia* sp., and *Candidatus* Jettenia asiatica, all of them represented in low relative abundance with the exception of *Candidatus Brocadia* sp. Sequences for the OTUs identified as *Candidatus Brocadia* sp. in both Lab MBR and Low Temperature CANON bioreactors (305 and 289 nucleotides, resp.) shared 99% similarity in 100% query cover with four *Candidatus Brocadia fulgida* sequences found in the BLASTN database (JQ864319.1, JQ864321.1, JQ864322.1, Zheng & Zhang, unpublished, and JX243455.1 [36]). Thus, it can be said that universal primer 530F can express in a consistent way the diversity of anammox bacteria found in autotrophic nitrogen removal bioreactors.

Failure to find *Candidatus* Brocadiales sequences at annealing temperature of 44°C can be related to a bias of PCR procedure. The relative abundance of *Brocadia* sp. and *Carboxydibrachium* sp. along with their ratio for the bioreactor Lab MBR can be seen in Table 2. The *Carboxydibrachium* sp./*Brocadia* sp. ratio abruptly increases when dropping from 45°C to 44°C, while it stabilizes at temperatures of 45°C and higher. In accordance with our results, some studies have found that a small percentage of OTUs in chicken caecal samples suffered from distortions in relative abundance as their bacterial community structure was studied through RT-PCR at different annealing temperatures, even though the entire microbial community structures were not subjected to major changes [37]. Also, changes in amplification of *Vibrio vulnificus* by RT-PCR at different annealing temperatures have been found [38]. Mismatches of primers with the targeted region of the DNA templates are thought to be the cause of differences in the estimation of the relative abundance of bacterial species in environmental samples. A higher number of mismatches in the targeted region lead to a higher bias of the relative abundance of certain OTUs. Among other factors, the annealing temperature is thought to cause mismatch of primers during PCR processes [33]. In this

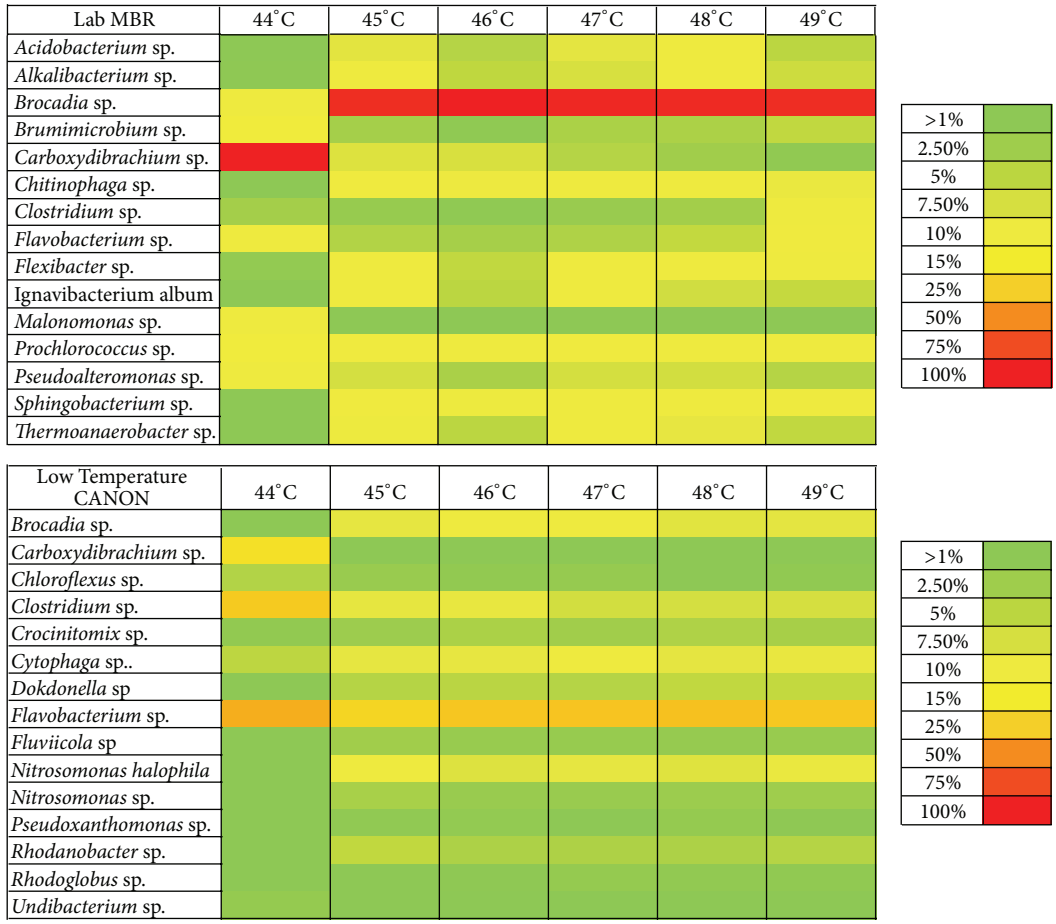


Figure 1: Heat maps of OTUs > 1% relative abundance in lab-scale bioreactors LAB MBR and Low Temperature CANON.

Table 2: Relative abundances of *Brocadia* sp., *Carboxydibrachium* sp. and the ratio of relative abundances *Carboxydibrachium* sp. to *Brocadia* sp. for the Lab MBR reactor.

	Lab MBR					
	44°C	45°C	46°C	47°C	48°C	49°C
Species						
Brocadia sp. (%)	1.04	75.15	86.76	79.51	77.68	75.77
Carboxydibrachium sp. (%)	84.50	0.81	0.77	0.44	0.26	0.13
C. sp./*B.* sp.	81.34	0.01	0.01	0.01	0.00	0.00

way, low temperatures with respect to optimum have been related to the proliferation of amplicons with mismatches [39, 40]. On the other hand, it has been shown that high annealing temperatures with respect to optimum tend to increase the ratio of amplification of one-mismatch to no-mismatch sequences [41, 42]. Regardless of the diversity of opinions, knowledge of the factors that intervene in the differences of relative abundance of OTUs at different annealing temperatures is still incomplete; hence, primer mismatches may not be the only cause driving these changes [37].

3.5. CA and PCoA. CA a) and b) are shown in Figure S2. When the similarity of all OTUs identified was compared, there were no relevant differences between results generated by PCR at the different annealing temperatures assayed (CA a)). However, CA b), which analyzes only OTUs with >1% relative abundance, clearly shows significant differences between the results generated at 44°C compared to the rest of annealing temperatures tested. This is in accordance with the heat maps generated for the bioreactors sampled (Figure 1). This implies that Bray-Curtis dissimilarities between annealing temperatures are less pronounced when samples are studied to their full sampling depths.

Phylogeny-dependent CA c) for each bioreactor is shown in Figure S3. Similarity between samples varies depending on the bioreactor analyzed. Samples from lab-scale bioreactors showed a higher phylogeny-based similarity for

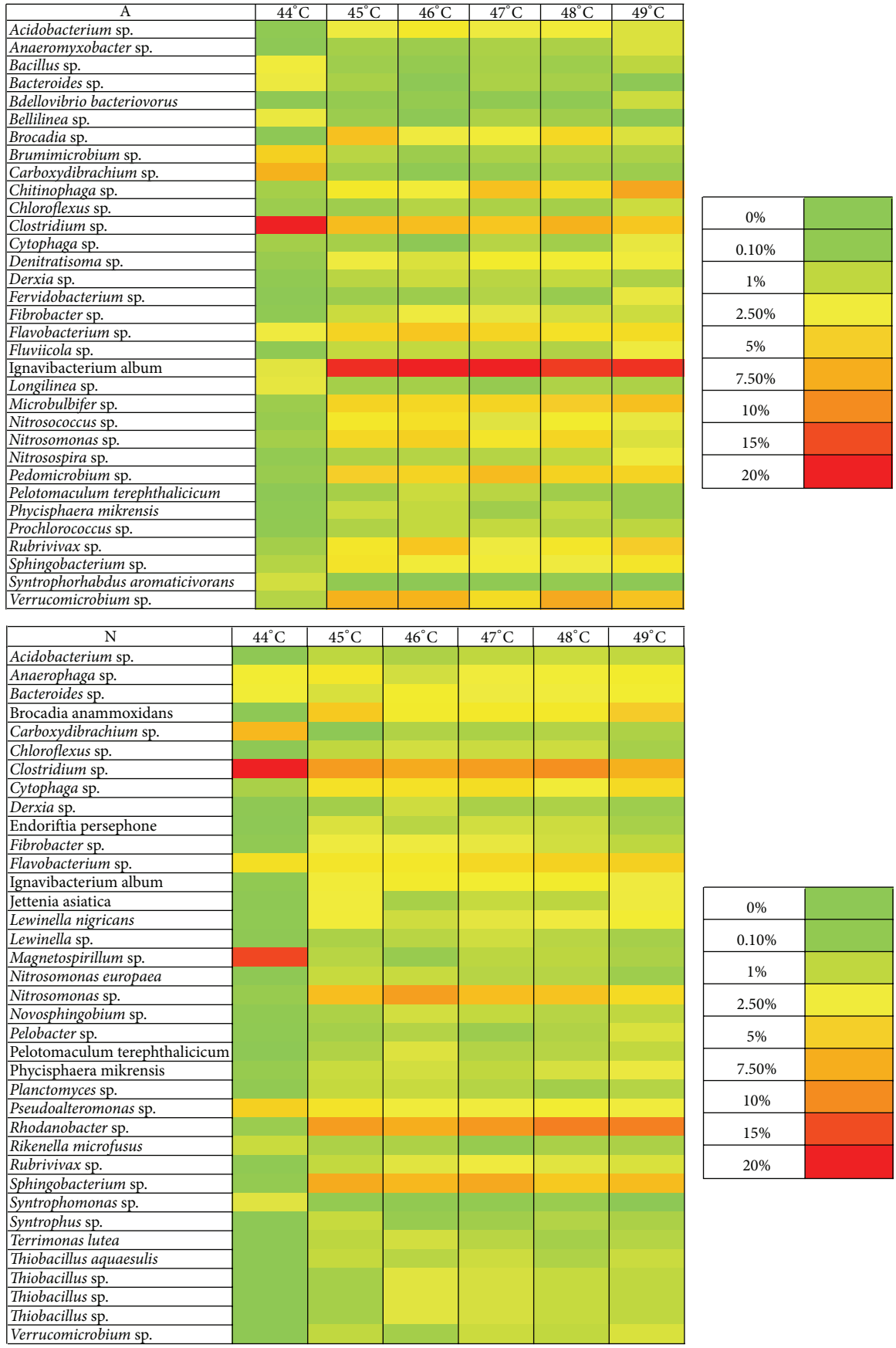


Figure 2: Continued.

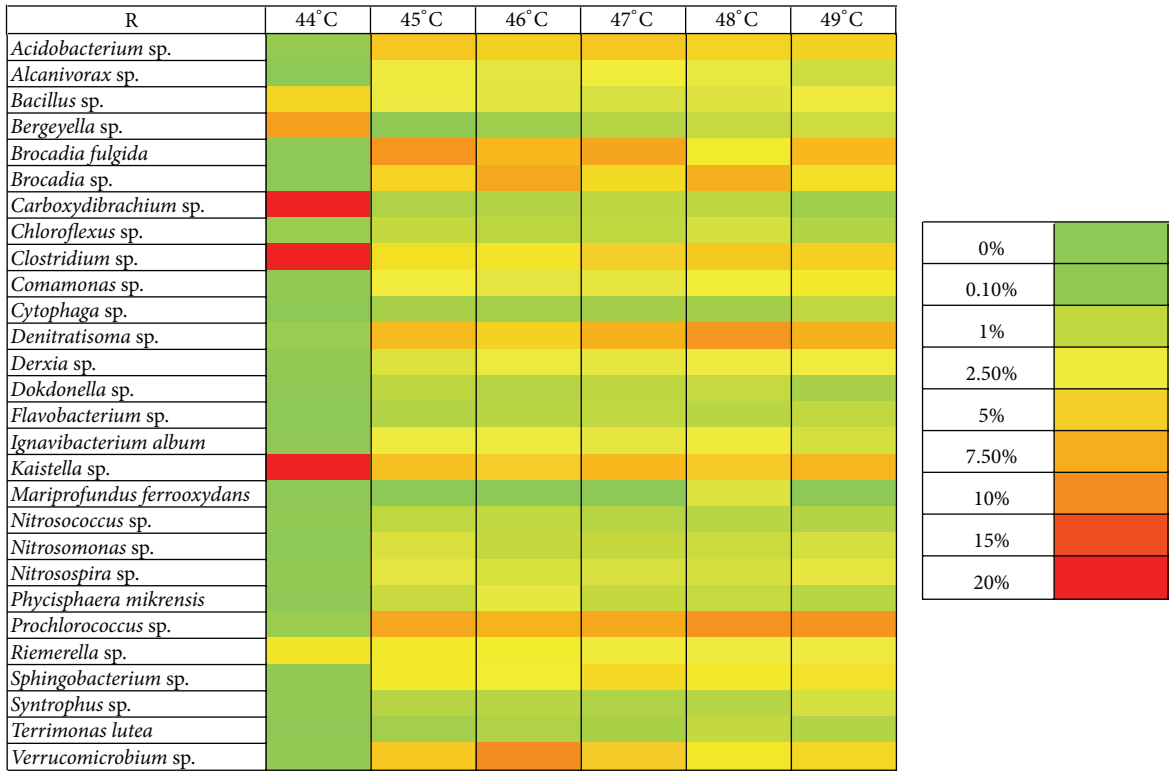


Figure 2: Heat maps of OTUs > 1% relative abundance in full scale bioreactors Appeldorn (A), Olburgen (N), and Rotterdam (R).

all the annealing temperatures tested. No significant differences at 80% similarity were found in bioreactor Low Temperature CANON, and three clusters were formed at 90% similarity. For bioreactor Lab MBR there were no significant differences at 70% similarity, but three clusters were generated at 80% and 90% similarity. Full-scale bioreactors showed lower phylogeny-based similarity, with significant differences at 70% for bioreactors N and R and sequences grouped in five or six clusters at 90% similarity in all cases. This leads to the assumption that differences in bacterial communities obtained at different annealing temperatures are more pronounced in more complex ecosystems.

Phylogeny-dependent PCoA 2D views are offered in Figure S4. PCoA was done to double check the results obtained by CA of samples. Phylogeny-dependent PCoA showed that 44°C tended to discriminate from the other annealing temperatures in terms of microbial community structure explanation. On the other hand, samples annealed at 45°C and 48°C seemed to be related in terms of microbial community structure. As also proven by cluster analysis, the differences between various annealing temperatures were found more pronounced as the complexity of the bacterial communities increased.

Some authors have reported that differences in annealing temperature do not significantly affect the microbial community structure of chicken caecal samples [37]. In our case, differences in microbial community structure at both full sampling depth and at >1% OTUs were confirmed.

3.6. Species Richness and Hill Diversity Indices. Values for individuals computed and original, interpolated and extrapolated species richness are summarized in Table 3. Rating for performance of the different annealing temperatures based on species richness is shown in Table 4. Differences in species richness are accounted at the original sequencing depth, interpolated to the lowest number of reads per bioreactor, and extrapolated to the highest number of reads per bioreactor. Even though species richness comparison cannot be performed at different sequencing depths, methods for interpolation and extrapolation of rarefaction curves can precisely estimate species richness diversity of samples with different number of reads, making comparison between them possible [29].

Differences in species richness and number of reads obtained at each annealing temperature can be appreciated. As a main trend, 44°C is the temperature that offered the highest number of reads, followed by 45°C and 46°C. It has been suggested that annealing temperatures lower than the optimum increase the number of PCR products generated [43]. The lowest number of reads was obtained at 49°C. 48°C was the temperature generating a higher interpolated species richness, while at 44°C a poorer performance was observed, and 44°C was also the worst among all annealing temperatures regarding original species richness. Extrapolated species richness of samples was higher at 45°C, followed by 48°C.

In conclusion, 45°C is the annealing temperature showing the best performance regarding the expression of species

Table 3: Values for individuals computed, original species richness *S*(est), interpolated species richness *S*(int), and extrapolated species richness *S*(ext).

	Temperature	Individuals (computed)	*S*(est)	*S*(int)	*S*(ext)
Lab MBR (T2)	44°C	11646	42	34.8	42.99
	45°C	16895	148	85.6	148.4
	46°C	17336	117	67	117
	47°C	3162	78	76.8	85.67
	48°C	5883	100	79.2	108.9
	49°C	3024	105	105	117.3
Low Temperature CANON (TC)	44°C	11438	61	30.1	38.65
	45°C	14362	179	100.3	179
	46°C	9545	155	94.4	163.1
	47°C	4145	119	99.5	131.6
	48°C	5068	133	99.3	147.9
	49°C	2394	87	87	95.53
Apeldoorn DEMON (A)	44°C	9226	78	29.3	78.76
	45°C	9924	161.7	45.1	161.7
	46°C	654	73	55.6	79.11
	47°C	2192	141	66.5	156.9
	48°C	8206	258	68	265.2
	49°C	321	60	59.9	68.55
Olburgen CANON (N)	44°C	17193	87	41.9	87
	45°C	2893	198	134.2	218.3
	46°C	947	141	140.9	160.7
	47°C	3180	205	135.7	221.9
	48°C	2462	197	139.2	217.1
	49°C	1905	174	133.8	192.6
Rotterdam 2-stage anammox (R)	44°C	12977	81	43.2	81
	45°C	5183	211	147.3	228
	46°C	3879	188	143.5	205.2
	47°C	4120	198	151.6	212.3
	48°C	1679	143	143	159.6
	49°C	1791	145	142.3	159.7

Table 4: Ratings for the comparison of species richness for samples from different bioreactors at each annealing temperature. Higher mean ratio is related to higher performance of the annealing temperature.

Temperature	Mean ratio individuals	Mean ratio *S*(int)	Mean ratio *S*(ext)
44°C	89.36%	32.89%	31.00%
45°C	72.36%	88.05%	91.86%
46°C	41.79%	86.87%	72.43%
47°C	24.22%	93.29%	76.70%
48°C	37.09%	93.51%	84.76%
49°C	12.49%	92.73%	63.01%

richness of bacterial communities in the bioreactors studied, while annealing at 49°C offers a worse performance taking into account the low number of reads consistently obtained.

Hill diversity indices of first order (Shannon index) and second order (Simpson index) have been defended as the most robust method for comparison of bacterial assemblages from natural ecosystems [31]. Hill diversities of first order and second order for each sample are shown in Table 5. Ratings for the comparison of each annealing temperature based on the indices can be seen in Table 6. Both indices showed the same quality pattern, with 45°C being the highest value followed closely by 49°C and 48°C. Once again, 45°C

Table 5: Values for the Shannon index and the Simpson index of all samples analyzed in the study.

	Temperature	Shannon index	Simpson index
Lab MBR (T2)	44°C	1.559179	0.4866806
	45°C	1.657458	0.5265754
	46°C	0.9898754	0.2964455
	47°C	1.398619	0.4398892
	48°C	1.536843	0.480206
	49°C	1.665004	0.4945622
Low Temperature CANON (TC)	44°C	2.02004	0.7847361
	45°C	3.018484	0.9061173
	46°C	2.85764	0.8814217
	47°C	2.86951	0.8790168
	48°C	2.860865	0.8741682
	49°C	2.871034	0.8860556
Apeldoorn DEMON (A)	44°C	2.297351	0.6694541
	45°C	3.561685	0.9399574
	46°C	3.286833	0.928345
	47°C	3.51128	0.9295173
	48°C	3.651016	0.944623
	49°C	3.366082	0.9407517
Olburgen CANON (N)	44°C	2.959167	0.8714149
	45°C	3.979032	0.9632393
	46°C	4.039	0.9667856
	47°C	4.021377	0.9636953
	48°C	4.015465	0.9615441
	49°C	3.965338	0.9629424
Rotterdam 2-stage anammox (R)	44°C	2.497644	0.8499261
	45°C	3.910746	0.9644197
	46°C	3.87392	0.9619344
	47°C	3.903737	0.9636846
	48°C	3.865236	0.9630658
	49°C	3.902146	0.9643079

Table 6: Rating for the comparison of species diversity for samples from different bioreactors at each annealing temperature. Higher mean ratio is related to higher performance of the annealing temperature.

Temperature	Mean ratio Shannon index	Mean ratio Simpson index
44°C	72.12%	85.63%
45°C	99.12%	99.83%
46°C	88.64%	90.32%
47°C	94.92%	95.71%
48°C	97.07%	97.40%
49°C	97.14%	98.24%

was the best annealing temperature for capturing diversity of bacterial communities inside the bioreactors analyzed.

No significant differences in species richness and evenness of chicken caecal samples studied through RT-PCR have been reported [37]. Nevertheless, other authors have shown that differences in annealing temperature lead to different bonding of primer and targeted region of genetic templates, having an impact over ecological parameters of environmental samples. This is caused by enhanced amplification of certain strains, which deviates relative abundance of these species, or by nonspecific bonding, which increases the microbial diversity recorded on the samples [39, 44]. In our case, small differences in annealing temperature changed species richness, effective number of counts, and Hill

diversities within the samples. Samples processed at 44°C accounted for the highest number of reads but surprisingly also contained the lowest species richness and diversity. This shows that annealing temperatures of 44°C reduced the estimation of the diversity of the system. Reduced diversity of samples analyzed at 44°C may rest on the fact that the primer is subjected to much more stringent bonding than that at the other annealing temperatures, therefore generating many sequences of a low number of species. All other annealing temperatures can capture the diversity of the ecosystem studied in a consistent fashion. 46°C offers a lower diversity quality values in comparison with the others.

In terms of species richness and diversity, 44°C offers poor results compared to the others, but it can be said that annealing temperatures ranging from 45 to 49°C offer good results. Nevertheless, 45°C stands as the optimal annealing temperature of all those tested due to superior species richness and diversity indices values. Interestingly, it yields better results in species richness and diversity than the *in silico* calculated salt-adjusted optimum for the primer. As suggested by Hecker and Roux, 1996 [43], annealing temperatures above and down the optimum annealing temperature theoretically calculated for a given primer increase number of reads obtained and specificity of PCR products. In this case, annealing temperature at 45°C gives better results, in terms of species richness and diversity, than the optimum salt-adjusted annealing temperature for the primer utilized.

4. Conclusions

The superior coverage of primer 530F with respect to other popular universal primers was indicated through *in silico* testing. The ability of primer 530F to target the majority of known anammox phylotypes was also demonstrated by *in silico* testing. Therefore, primer 530F stands as the best universal primer available for the metagenomic analysis of microbial communities where anammox bacteria are expected to develop important ecological functions.

The 30 pyrosequencing analysis showed that the annealing temperature produces a severe effect over the microbial community structure discerned through pyrosequencing. Strong bias in the identification of anammox species in particular at annealing temperature of 44°C was observed in all the pyrosequencing samples for each of the five autotrophic nitrogen removal technologies. The six different annealing temperatures analyzed in the study showed different microbial community structure compositions as proved by phylogeny-based and non-phylogeny-based CA and PCoA. Differences among annealing temperatures could also be observed with respect to the number of individuals sequenced and species richness with 45°C being the best in these terms. In this sense, annealing temperatures of 45°C demonstrate good coverage of total bacteria and anammox species, high number of reads, the highest species richness, and the highest diversity indices values. Therefore, results show that autotrophic nitrogen removal bioreactor bacterial community analyses including anammox bacteria can be done using the universal primer 530F with annealing temperature of 45°C. Under these conditions, this procedure is a good approach for the bacterial diversity study in autotrophic nitrogen removal technologies using pyrosequencing methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Alejandro Gonzalez-Martinez was responsible for the bioreactors field study and *in silico* testing of the different primers. Alejandro Gonzalez-Martinez was mainly responsible for molecular biology experimental procedures, and Ben A. Abbas was also responsible for this. Alejandro Rodriguez-Sanchez was mainly responsible for the statistical analysis of samples and microbial ecology analysis, while Maria Victoria Martinez-Toledo and Belén Rodelas also were responsible for microbial ecology analysis. Alejandro Gonzalez-Martinez, Alejandro Rodriguez-Sanchez, Belén Rodelas, and Jesus Gonzalez-Lopez were responsible for writing the paper. Alejandro Gonzalez-Martinez and Alejandro Rodriguez-Sanchez were responsible for the conformation of all tables and figures. Alejandro Gonzalez-Martinez, Belén Rodelas, Maria Victoria Martinez-Toledo, Mark C. M. van Loosdrecht, F. Osorio, and Jesus Gonzalez-Lopez were responsible for revision of the paper.

Acknowledgement

The authors received a financial support from a collaborator company but it has not participated in the work development.

References

[1] M. Strous, J. A. Fuerst, E. H. M. Kramer et al., "Missing lithotroph identified as new planctomycete," *Nature*, vol. 400, no. 6743, pp. 446–449, 1999.

[2] T. Dalsgaard, B. Thamdrup, and D. E. Canfield, "Anaerobic ammonium oxidation (anammox) in the marine environment," *Research in Microbiology*, vol. 156, no. 4, pp. 457–464, 2005.

[3] M. M. M. Kuypers, A. O. Silekers, G. Lavik et al., "Anaerobic ammonium oxidation by anammox bacteria in the Black Sea," *Nature*, vol. 422, no. 6932, pp. 608–611, 2003.

[4] B. Thamdrup, T. Dalsgaard, M. M. Jensen, O. Ulloa, L. Farías, and R. Escribano, "Anaerobic ammonium oxidation in the oxygen-deficient waters off Northern Chile," *Limnology and Oceanography*, vol. 51, no. 5, pp. 2145–2156, 2006.

[5] S. Ganesh, D. J. Parris, E. F. Delong, and F. J. Stewart, "Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone," *ISME Journal*, vol. 8, no. 1, pp. 187–211, 2014.

[6] L.-D. Shen, S. Liu, L.-P. Lou et al., "Broad distribution of diverse anaerobic ammonium-oxidizing bacteria in Chinese agricultural soils," *Applied and Environmental Microbiology*, vol. 79, no. 19, pp. 6167–6172, 2013.

[7] S. M. Kotay, B. L. Mansell, M. Hogsett, H. Pei, and R. Goel, "Anaerobic ammonia oxidation (ANAMMOX) for side-stream treatment of anaerobic digester filtrate process performance and

microbiology," *Biotechnology and Bioengineering*, vol. 110, no. 4, pp. 1180–1192, 2013.

[8] A. González-Martínez, J. M. Poyatos, E. Hontoria, J. Gonzalez-Lopez, and F. Osorio, "Treatment of effluents polluted by nitrogen with new biological technologies based on autotrophic nitrification-denitrification processes," *Recent Patents on Biotechnology*, vol. 5, no. 2, pp. 74–84, 2011.

[9] M. S. M. Jetten, M. Wagner, J. Fuerst, M. C. M. van Loosdrecht, G. Kuenen, and M. Strous, "Microbiology and application of the anaerobic ammonium oxidation ('anammox') process," *Current Opinion in Biotechnology*, vol. 12, no. 3, pp. 283–288, 2001.

[10] S. Sri Shalini and K. Joseph, "Nitrogen management in landfill leachate: application of SHARON, ANAMMOX and combined SHARON-ANAMMOX process," *Waste Management*, vol. 32, no. 12, pp. 2385–2400, 2012.

[11] T. Liu, D. Li, H. Zeng et al., "Biodiversity and quantification of functional bacteria in completely autotrophic nitrogen-removal over nitrite (CANON) process," *Bioresource Technology*, vol. 118, pp. 399–406, 2012.

[12] Z.-X. Quan, S.-K. Rhee, J.-E. Zuo et al., "Diversity of ammonium-oxidizing bacteria in a granular sludge anaerobic ammonium-oxidizing (anammox) reactor," *Environmental Microbiology*, vol. 10, no. 11, pp. 3130–3139, 2008.

[13] Z. Hu, D. R. Speth, K.-J. Francoijs, Z.-X. Quan, and M. S. M. Jetten, "Metagenome analysis of a complex community reveals the metabolic blueprint of anammox bacterium '*Candidatus Jettenia asiatica*'," *Frontiers in Microbiology*, vol. 3, article 366, 2012.

[14] Q. Yang, Z. Jia, R. Liu, and J. Chen, "Molecular diversity and anammox activity of novel planctomycete-like bacteria in the wastewater treatment system of a full-scale alcohol manufacturing plant," *Process Biochemistry*, vol. 42, no. 2, pp. 180–187, 2007.

[15] X.-R. Li, B. Du, H.-X. Fu et al., "The bacterial diversity in an anaerobic ammonium-oxidizing (anammox) reactor community," *Systematic and Applied Microbiology*, vol. 32, no. 4, pp. 278–289, 2009.

[16] B.-J. Ni, B.-L. Hu, F. Fang et al., "Microbial and physicochemical characteristics of compact anaerobic ammonium-oxidizing granules in an upflow anaerobic sludge blanket reactor," *Applied and Environmental Microbiology*, vol. 76, no. 8, pp. 2652–2656, 2010.

[17] P. Han, Y.-T. Huang, J.-G. Lin, and J.-D. Gu, "A comparison of two 16S rRNA gene-based PCR primer sets in unraveling anammox bacteria from different environmental samples," *Applied Microbiology and Biotechnology*, vol. 97, no. 24, pp. 10521–10529, 2013.

[18] B. Xie, Z. Lv, C. Hu, X. Yang, and X. Li, "Nitrogen removal through different pathways in an aged refuse bioreactor treating mature landfill leachate," *Applied Microbiology and Biotechnology*, vol. 97, no. 20, pp. 9225–9234, 2013.

[19] M. S. M. Jetten, L. van Niftrik, M. Strous, B. Kartal, J. T. Keltjens, and H. J. M. Op Den Camp, "Biochemistry and molecular biology of anammox bacteria," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 44, no. 2-3, pp. 65–84, 2009.

[20] P. Junier, V. Molina, C. Dorador et al., "Phylogenetic and functional marker genes to study ammonia-oxidizing microorganisms (AOM) in the environment," *Applied Microbiology and Biotechnology*, vol. 85, no. 3, pp. 425–440, 2010.

[21] S. E. Dowd, T. R. Callaway, R. D. Wolcott et al., "Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP)," *BMC Microbiology*, vol. 8, article 125, 2008.

[22] W. Rychlik, W. J. Spencer, and R. E. Rhoads, "Optimization of the annealing temperature for DNA amplification in vitro," *Nucleic Acids Research*, vol. 18, no. 21, pp. 6409–6412, 1990.

[23] X. Chen, P. Liu, and H.-H. Chou, "Whole-genome thermodynamic analysis reduces siRNA off-target effects," *PLoS ONE*, vol. 8, no. 3, Article ID e58326, 2013.

[24] D. Dobrijevic, G. Di Liberto, K. Tanaka et al., "High-throughput system for the presentation of secreted and surface-exposed proteins from Gram-positive bacteria in functional metagenomic studies," *PLoS ONE*, vol. 8, no. 6, Article ID e65956, 2013.

[25] S. Nakano, Y. Morizane, N. Makisaka et al., "Development of a highly sensitive immuno-PCR assay for the measurement of α-galactosidase A protein levels in serum and plasma," *PLoS ONE*, vol. 8, no. 11, Article ID e78588, 2013.

[26] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[27] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, "UCHIME improves sensitivity and speed of chimera detection," *Bioinformatics*, vol. 27, no. 16, pp. 2194–2200, 2011.

[28] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome Biology*, vol. 15, no. 3, article R46, 2014.

[29] R. K. Colwell, A. Chao, N. J. Gotelli et al., "Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages," *Journal of Plant Ecology*, vol. 5, no. 1, pp. 3–21, 2012.

[30] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, http://www.R-project.org.

[31] B. Haegeman, J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, and J. S. Weitz, "Robust estimation of microbial diversity in theory and in practice," *The ISME Journal*, vol. 7, no. 6, pp. 1092–1101, 2013.

[32] K. Malhotra, L. Foltz, W. C. Mahoney, and P. A. Schueler, "Interaction and effect of annealing temperature on primers used in differential display RT-PCR," *Nucleic Acids Research*, vol. 26, no. 3, pp. 854–856, 1998.

[33] D. Bru, F. Martin-Laurent, and L. Philippot, "Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example," *Applied and Environmental Microbiology*, vol. 74, no. 5, pp. 1660–1663, 2008.

[34] W. R. L. van der Star, A. I. Miclea, U. G. J. M. van Dongen, G. Muyzer, C. Picioreanu, and M. C. M. van Loosdrecht, "The membrane bioreactor: a novel tool to grow anammox bacteria as free cells," *Biotechnology and Bioengineering*, vol. 101, no. 2, pp. 286–294, 2008.

[35] T. Lotti, *Monitoraggio del proceso anammox: aspetti fisiologici per unutilizzo in piena scala*, Università degli Studi della Basilicata, Potenza, Italy, 2011.

[36] Z. Hu, T. Lotti, M. de Kreuk et al., "Nitrogen removal by a nitritation-anammox bioreactor at low temperature," *Applied and Environmental Microbiology*, vol. 79, no. 8, pp. 2807–2812, 2013.

[37] M. J. Sergeant, C. Constantinidou, T. Cogan, C. W. Penn, and M. J. Pallen, "High-throughput sequencing of 16s rRNA gene amplicons: Effects of extraction procedure, primer length and annealing temperature," *PLoS ONE*, vol. 7, no. 5, Article ID e38094, 2012.

[38] S. Wang and R. E. Levin, "Thermal factors influencing detection of *Vibrio vulnificus* using real-time PCR," *Journal of Microbiological Methods*, vol. 69, no. 2, pp. 358–363, 2007.

[39] J. T. Hsu, S. Das, and S. Mohapatra, "Polymerase chain reaction engineering," *Biotechnology and Bioengineering*, vol. 55, no. 2, pp. 359–366, 1997.

[40] M. J. Claesson, Q. Wang, O. O'Sullivan et al., "Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions," *Nucleic Acids Research*, vol. 38, article e200, 2010.

[41] K. Ishii and M. Fukui, "Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR," *Applied and Environmental Microbiology*, vol. 67, no. 8, pp. 3753–3755, 2001.

[42] R. Sipos, A. J. Székely, M. Palatinszky, S. Révész, K. Márialigeti, and M. Nikolausz, "Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis," *FEMS Microbiology Ecology*, vol. 60, no. 2, pp. 341–350, 2007.

[43] K. H. Hecker and K. H. Roux, "High and low annealing temperatures increase both specificity and yield in touchdown and stepdown PCR," *BioTechniques*, vol. 20, no. 3, pp. 478–485, 1996.

[44] K. Malhotra, L. Foltz, W. C. Mahoney, and P. A. Schueler, "Interaction and effect of annealing temperature on primers used in differential display RT-PCR," *Nucleic Acids Research*, vol. 26, no. 3, pp. 854–856, 1998.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 167249, 5 pages
http://dx.doi.org/10.1155/2015/167249

Research Article

mmnet: An R Package for Metagenomics Systems Biology Analysis

Yang Cao,[1] Xiaofei Zheng,[2] Fei Li,[1] and Xiaochen Bo[1]

[1]*Department of Biotechnology, Beijing Institute of Radiation Medicine, 27 Taiping Road, Haidian District, Beijing 100850, China*
[2]*Department of Biochemistry and Molecular Biology, Beijing Institute of Radical Medicine, 27 Taiping Road, Haidian District, Beijing 100850, China*

Correspondence should be addressed to Fei Li; pittacus@gmail.com and Xiaochen Bo; boxc@bmi.ac.cn

Received 8 October 2014; Revised 8 March 2015; Accepted 17 April 2015

Academic Editor: Yunlong Liu

The human microbiome plays important roles in human health and disease. Previous microbiome studies focused mainly on single pure species function and overlooked the interactions in the complex communities on system-level. A metagenomic approach introduced recently integrates metagenomic data with community-level metabolic network modeling, but no comprehensive tool was available for such kind of approaches. To facilitate these kinds of studies, we developed an R package, *mmnet*, to implement community-level metabolic network reconstruction. The package also implements a set of functions for automatic analysis pipeline construction including functional annotation of metagenomic reads, abundance estimation of enzymatic genes, community-level metabolic network reconstruction, and integrated network analysis. The result can be represented in an intuitive way and sent to Cytoscape for further exploration. The package has substantial potentials in metagenomic studies that focus on identifying system-level variations of human microbiome associated with disease.

1. Introduction

The human microbiome has been proved to play a key role in human health and disease. Various microorganisms species along with wide range of interactions among them structure the microbial communities as inherently complex ecosystems across different human body sites, such as gut, oral cavity, and skin [1, 2]. The microbes in our bodies are crucial for human life, having profound influence on human physiology and health [3–5]. Most of these microbes are still unculturable and uncharacterized [6]. Furthermore, dynamic shifts in microbial community structure will change the immunity and metabolic function and cause various diseases, including obesity, inflammatory bowel disease (IBD), and Crohn's disease [7–9].

Traditional culture-dependent methods are restricted by the small number of cultured species and often fail to describe the less abundant species [10]. To address this challenge, metagenomics (defined as environmental and community genomics), a culture-independent technology, has been developed and widely used for microbial community analysis [11]. Specifically, unlike the initial capillary sequence-based or PCR-based metagenomic approaches, high-throughput metagenomic approaches based on next generation sequencing (NGS) make metagenomic analysis more sensitive, broader, and cheaper, providing critical insights into microbe-host interaction in large scale [12, 13].

A key challenge of applying metagenomics to microbial community is metabolic network reconstruction from metagenomic data. Previous studies focused mainly on the "parts list" of the microbiome and overlooked the interactions in the complex communities on system-level [14, 15]. However, an integrated approach to reconstruct the metabolic network by integrating metagenomic data with genome-scale metabolic modeling was introduced recently [7]. This metagenomic system biology method serves the entire microbiome as a single superorganism and utilizes the computational systems biology and complex network theory, providing comprehensive systems-level understanding of the microbiome by integrating metagenomic data with genome-scale metabolic modeling. However, there is no comprehensive tool available for such kind of approaches.

Here, we present an open source R package named *mmnet* to implement community-level metabolic network reconstruction. Moreover, *mmnet* implements a set of functions to construct an automatic pipeline from functional annotation of metagenomic reads to integrated network analysis. The result can be represented in an intuitive way and sent to Cytoscape for further exploration. The source code is published under GNU GPLv2 license and is freely available on the Bioconductor project (http://www.bioconductor.org/packages/devel/bioc/html/mmnet.html).

2. Methods and Implementation

2.1. Methods

Metagenomic Sequence Reads Annotation. To assess the functional capacity of microbial community, metagenomic sequence reads need alignment to a database of known genes and can be achieved with several well-characterized functional databases including KEGG orthology (KO) [16] and COGs [17]. The MG-RAST [18], a stable, extensible, online metagenomics analysis platform, is well maintained and provides a RESTful web API (Application Programming Interface) http://api.metagenomics.anl.gov/api.html. In this package, we annotate metagenomic reads by calling the RESTful web API of MG-RAST with R package *Rcurl*. If the metagenomic sequences have been annotated on MG-RAST, it will return the corresponding metagenome ID that already exists in MG-RAST without duplicated annotation.

Enzymatic Gene Abundances Estimation. M5NR [19], MG-RAST uses to annotate sequences, is an integration of many sequence databases into one single comprehensive, searchable database including most common functional databases like KO, EBI, and SEED. Once metagenomic sequences were annotated, KO information was filtered and extracted in the annotation profile from MG-RAST for subsequent metabolic network reconstruction. Enzymatic genes abundances can be estimated according to the following guidelines. (1) The count for the sequence matched a single reference sequence that had been annotated with more than one KO split evenly between all the KO annotations. (2) The count for the sequence matched more than one KO-annotated reference sequence with the same *e*-value split evenly between KOs of all matched reference sequences. Relative abundances of enzymatic genes in each sample can be finally computed by normalizing the counts of the reads for each KO with the total number of reads of enzymatic genes accounting for the different sample depths.

System-Level Metabolic Network Construction. A system-level metabolic network was constructed from the entire enzymatic genes (KO) in any sample. Reference metabolic data used to construct metabolic network, which consists of KO id and metabolic reactions, was obtained by phasing KEGG metabolic pathway dataset using KEGG REST API. Each enzyme may be associated with multiple reactions, and each reaction may be associated with multiple enzymes. In this metabolic network, enzymes are connected with directed edges, and a directed edge from enzyme A to enzyme B indicates that a product metabolite of a reaction catalyzed by enzyme A is a substrate metabolite of a reaction catalyzed by enzyme B.

Table 1: The topological properties calculated in the SSN.

Topological features	Description
Betweenness centrality	The fraction of shortest paths between node pairs that pass through the node
Clustering coefficient	The number of triangles (3 loops) that pass through this node
PageRank	The number and PageRank metric of all nodes that link to the node
Degree	The number of edges connected to the node

To examine whether enzymes that are associated with a specific host state exhibit some topological features in the SSN, we calculated most common topological properties (Table 1) of each node in the network.

Metabolic Network Analysis. Comparing the abundances of enzymatic genes in different samples (e.g., disease and healthy samples) can reveal enzymes associated with specific host state. The package mmnet implements three methods to measure the differential abundance between different samples, including odds ratio (OR), RANK, and Jensen-Shannon Diverge (JSD). First, OR was calculated according to

$$\mathrm{OR}(k) = \frac{\left[\sum_{s=\mathrm{state}_1} A_{sk} / \sum_{s=\mathrm{state}_1} \left(\sum_{i\neq k} A_{si}\right)\right]}{\left[\sum_{s=\mathrm{state}_2} A_{sk} / \sum_{s=\mathrm{state}_2} \left(\sum_{i\neq k} A_{si}\right)\right]}, \quad (1)$$

where A_{sk} represents the abundance of enzyme k in sample s and state_1 and state_2 represent two types of samples with different state. The differential abundance score was defined as the absolute value of the fold change in OR, $\mathrm{abs}[\log_2(\mathrm{OR})]$. Once method RANK was selected, the enzyme abundances were ranked within each sample from most abundant to least abundant first. The difference abundance score was then measured with the difference between the mean ranks of samples in different state. Finally, users can examine the divergence using the Jensen-Shannon divergence algorithm to quantify the differential abundance score between samples in different state.

2.2. Implementation.

A typical analysis pipeline for metagenomic systems biology supported by this package (Figure 1) starts with functional annotation and abundance estimation, then state specific network (SSN) construction, and finally topological and differential network analysis. *mmnet* is released as an R package including seven main functions (Figure 1): *constructSSN, submitMgrastJob, estimateAbundance, updateKEGGPathway, constructMetabolicNetwork, topologicalAnalyzeNet, and differentialAnalyzeNet*. All of these functions will be introduced as follows.

2.2.1. Metagenomic Sequence Reads Annotation. submitMgrastJob is used for metagenomic sequence reads annotation

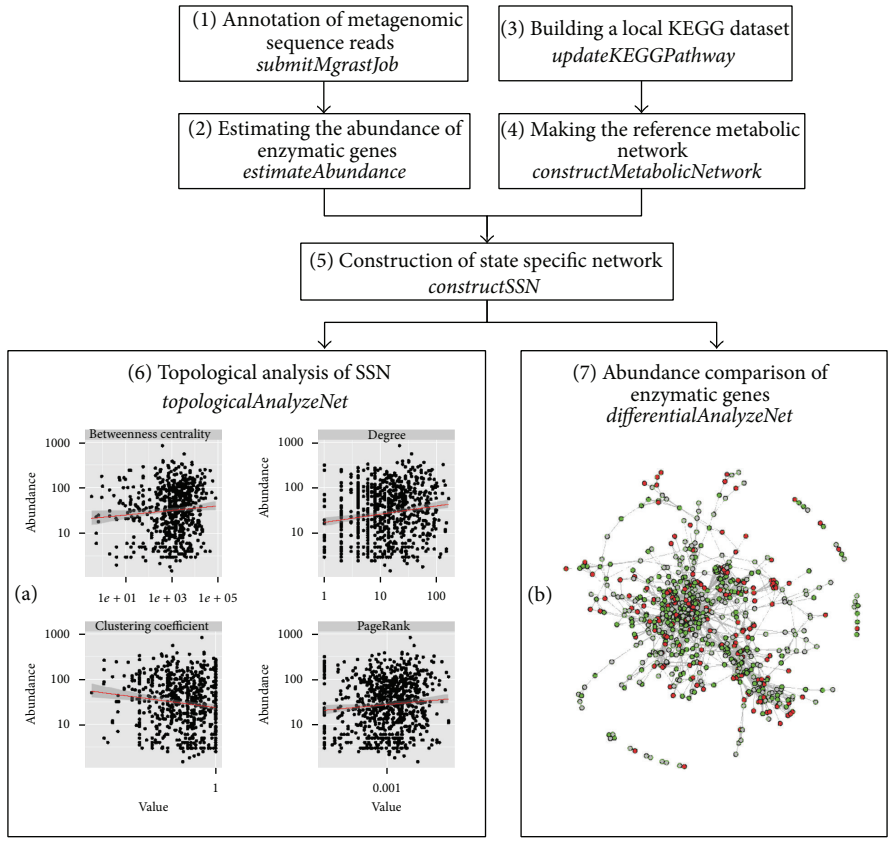


Figure 1: A typical analysis pipeline supported by the mmnet package.

on MG-RAST. Before sequence annotation, users should log in into MG-RAST first. Acceptable sequence data can be in FASTA, FASTQ, or SFF format. A MG-RAST ID represents that the annotation data is returned by this function. Then annotation profile will be obtained using function *getMgrastAnnotation* which takes the MG-RAT ID as input. The output is a data-frame in which one row corresponds to one KO annotation information.

2.2.2. Estimating the Abundance of Enzymatic Genes. estimateAbundance is to estimate the abundances of enzymatic genes in the annotation profile mentioned above. As a result, a Biological Observation Matrix (BIOM) [20] file format, which is designed to be a general-use format for representing biological sample by observation contingency tables, is returned to encode enzymatic gene abundance profiles.

2.2.3. Building a Local KEGG Dataset. updateKEGGPathway is used to build or update a local KEGG metabolic dataset for improving the efficiency of analysis and avoiding frequently repeated data downloads. This function is provided in *mmnet* package for users to build and update a local version of KEGG metabolic pathway dataset. The local metabolic data named *RefDbcache* is saved in the "\.*mmnet*" subdirectory under user-specified folder, default under the user's home directory.

2.2.4. Making the Reference Metabolic Network. constructMetabolicNetwork is used for construction of the reference metabolic network based on the metabolic reaction relationships in KEGG metabolic pathway dataset. A prebuilt reference metabolic network of class *igraph* has been integrated in *mmnet* package, so that the rebuilding of the network is unnecessary unless there are updates on the KEGG metabolic pathways.

2.2.5. Construction of STATE Specific Networks. constructSSN is designed to construct a SSN by calling function *constructMetabolicNetwork*. Its input is an abundance profile of a sample and the output is subnetworks of the reference network composed of only the enzymatic genes identified from samples in a given biological state. This network is also of class *igraph*. In addition, the abundances of enzymatic genes are taken as node attributes in the SSN. The resulting SSNs can be seamlessly analyzed in R environment by built-in functions or in Cytoscape [21] by utilizing RCytoscape package [22].

2.2.6. Topological Analysis of SSN. topologicalAnalyzeNet is to compute and illustrate the correlations between the topological properties of enzymatic genes and their abundances (Figure 1(a)). The input of the function is a single SSN. The output is also an *igraph*, in which all the calculated topological

properties of nodes in SSN are stored as node's attributes and can be exported for further analysis.

2.2.7. Differential Analysis of SSN. *Differential AnalyzeNet* is to calculate the differential abundance score of metabolic networks with different states. The function takes a list of SSNs as input and outputs a community-level metabolic network in which the significantly enriched or depleted enzymes are marked by colors (Figure 1(b)).

More detailed description of these functions and the package instructions is referred to in reference manual http://www.bioconductor.org/packages/devel/bioc/vignettes/mmnet/inst/doc/mmnet.pdf.

3. Results

To illustrate the analysis pipeline made by this package, we use a part of the public dataset containing 18 microbiomes from 6 obese and lean monozygotic twin pairs and their mothers [8]. 454 FLX pyrosequencer was used to carry out deep metagenomic shotgun sequencing of total fecal community DNA of 18 obese or lean samples. These metagenomic sequences have been annotated on MG-RAST, and the annotated data is available in MG-RAST. Thus, we downloaded the annotation profiles of these samples using *getMgrastAnnotation* directly without submitting a MG-RAST job. Apparently, users can annotate the sequenced metagenomic reads using function *submitMgrastJob* manually. For example, one of twin pairs (mgm4440616.3 and mgm4440824.3) and the function annotations can be accessed as follows:

```
source("http://bioconductor.org/biocLite
.R")
biocLite("mmnet")
library(mmnet)
pid <- c("4440616.3","4440824.3")
names(pid) <- c("obese", "lean")
annot <- lapply(pid,
getMgrastAnnotation)
```

The relative abundances of enzymatic genes in the two samples were estimated from the functional annotations, and then the corresponding SSNs were built. For these two samples, 1345 KOs were identified in total. The correlation coefficient tested with Pearson's method is 0.92, which indicated that the relative enzymatic gene abundance across these two samples was highly concordant:

```
abund <- estimateAbundance(annot)
ssn <- constructSSN(abund)
```

Based on the SSNs, we intuitively explored the correlations between the topological features of enzyme in the SSN and their abundance (Figure 1(a)) and performed differential network analysis to identify potential enzymes associated with obese (Figure 1(b)). After differential metabolic network analysis, enzymatic genes that are associated with specific state appear as colored nodes (red = enriched and green = depleted):

```
lapply(ssn, topologicalAnalyzeNet)
differentialAnalyzeNet(ssn, sample.
state = names(pid))
```

Notably, only two samples were taken for testing *mmnet* accounting for saving computing resources and time; the results will be more meaningful when more samples were tooken for analysis.

4. Conclusions

The metagenomic approach on metabolic network provides a system-level understanding of the microbiome and gains insight into variation in metabolic capacity. It is very useful for studying the metabolic activity and specifically complex inherent interactions by serving the microbial community as a single supraorganism. In this paper, we present the *mmnet* package to support metagenomic network reconstruction as an integrated way in R environment and to build automatic pipelines running from metagenomic sequencing reads to community-level metabolic network. This package has substantial potentials for community metabolism analysis.

Availability and Requirements

project name: mmnet;

project home page: http://www.bioconductor.org/packages/devel/bioc/html/mmnet.html;

operating system(s): platform independent;

programming language: R;

other requirements: R 3.1.0 or higher;

license: GNU GPLv2;

any restrictions to use by nonacademics: none.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

Yang Cao implemented the package and wrote the user manual. Fei Li designed the structure and interface of the software and drafted the paper. Xiaofei Zheng, Xiaochen Bo participated in the design of the package and helped to draft the paper. All authors read and approved the final paper.

Acknowledgments

National Major Science and Technology Special Projects for New Drugs (2013ZX09304101), National Major Science and Technology Special Projects for Infectious Diseases (2013ZX10004216), National Key Technologies R&D Program for New Drugs (2012ZX09301-003), National Science

& Technology Pillar Program of China (2012BAI29B07), and National Nature Science Foundation of China (81102419). The authors thank Dr. Folker Meyer for kind response about the usages of MG-RAST API.

References

[1] The Human Microbiome Project Consortium, "A framework for human microbiome research," *Nature*, vol. 486, no. 7402, pp. 215–221, 2012.

[2] T. H. M. P. Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[3] Y. Soen, "Environmental disruption of host-microbe co-adaptation as a potential driving force in evolution," *Frontiers in Genetics*, vol. 5, article 168, 2014.

[4] J. C. Clemente, L. K. Ursell, L. W. Parfrey, and R. Knight, "The impact of the gut microbiota on human health: an integrative view," *Cell*, vol. 148, no. 6, pp. 1258–1270, 2012.

[5] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-bacterial mutualism in the human intestine," *Science*, vol. 307, no. 5717, pp. 1915–1920, 2005.

[6] R. I. Amann, W. Ludwig, and K.-H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation," *Microbiological Reviews*, vol. 59, no. 1, pp. 143–169, 1995.

[7] S. Greenblum, P. J. Turnbaugh, and E. Borenstein, "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 2, pp. 594–599, 2012.

[8] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, "An obesity-associated gut microbiome with increased capacity for energy harvest," *Nature*, vol. 444, no. 7122, pp. 1027–1031, 2006.

[9] C. Manichanh, L. Rigottier-Gois, E. Bonnaud et al., "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach," *Gut*, vol. 55, no. 2, pp. 205–211, 2006.

[10] I. C. A. Lacerda, F. C. O. Gomes, B. M. Borelli et al., "Identification of the bacterial community responsible for traditional fermentation during sour cassava starch, cachaça and minas cheese production using cultureindependent 16s rrna gene sequence analysis," *Brazilian Journal of Microbiology*, vol. 42, no. 2, pp. 650–657, 2011.

[11] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products," *Chemistry & Biology*, vol. 5, no. 10, pp. R245–R249, 1998.

[12] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 260–270, 2012.

[13] J. Handelsman, "Metagenomics: application of genomics to uncultured microorganisms," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 4, pp. 669–685, 2004.

[14] E. Borenstein, "Computational systems biology and in silico modeling of the human microbiome," *Briefings in Bioinformatics*, vol. 13, no. 6, pp. 769–780, 2012.

[15] R. Levy and E. Borenstein, "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 31, pp. 12804–12809, 2013.

[16] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[17] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, no. 5338, pp. 631–637, 1997.

[18] F. Meyer, D. Paarmann, M. D'Souza et al., "The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, article 386, 2008.

[19] A. Wilke, T. Harrison, J. Wilkening et al., "The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools," *BMC Bioinformatics*, vol. 13, no. 1, article 141, 2012.

[20] D. McDonald, J. C. Clemente, J. Kuczynski et al., "The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome," *GigaScience*, vol. 1, no. 1, p. 7, 2012.

[21] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[22] P. T. Shannon, M. Grimes, B. Kutlu, J. J. Bot, and D. J. Galas, "RCytoscape: rools for exploratory network analysis," *BMC Bioinformatics*, vol. 14, no. 1, article 217, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 647389, 11 pages
http://dx.doi.org/10.1155/2015/647389

Research Article

Genetic Interactions Explain Variance in Cingulate Amyloid Burden: An AV-45 PET Genome-Wide Association and Interaction Study in the ADNI Cohort

Jin Li,[1,2] Qiushi Zhang,[1,2,3] Feng Chen,[1] Jingwen Yan,[4,5,6] Sungeun Kim,[4] Lei Wang,[1,2] Weixing Feng,[1,2] Andrew J. Saykin,[4] Hong Liang,[1,2,4] and Li Shen[4,5,6]

[1] *Institute of Biomedical Engineering, College of Automation, Harbin Engineering University, 145 Nantong Street, Harbin 150001, China*
[2] *Center for Bioinformatics, College of Automation, Harbin Engineering University, 145 Nantong Street, Harbin 150001, China*
[3] *College of Information Engineering, Northeast Dianli University, 169 Changchun Street, Jilin 132012, China*
[4] *Center for Neuroimaging, Department of Radiology and Imaging Sciences, Indiana University School of Medicine, 355 West 16th Street, Suite 4100, Indianapolis, IN 46202, USA*
[5] *Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 West 10th Street, Suite 5000, Indianapolis, IN 46202, USA*
[6] *Department of Biohealth Informatics, Indiana University School of Informatics and Computing, Indianapolis, IN 46202, USA*

Correspondence should be addressed to Hong Liang; liangh@iu.edu and Li Shen; shenli@iu.edu

Received 15 December 2014; Accepted 17 March 2015

Academic Editor: Dongchun Liang

Alzheimer's disease (AD) is the most common neurodegenerative disorder. Using discrete disease status as the phenotype and computing statistics at the single marker level may not be able to address the underlying biological interactions that contribute to disease mechanism and may contribute to the issue of "missing heritability." We performed a genome-wide association study (GWAS) and a genome-wide interaction study (GWIS) of an amyloid imaging phenotype, using the data from Alzheimer's Disease Neuroimaging Initiative. We investigated the genetic main effects and interaction effects on cingulate amyloid-beta (Aβ) load in an effort to better understand the genetic etiology of Aβ deposition that is a widely studied AD biomarker. PLINK was used in the single marker GWAS, and INTERSNP was used to perform the two-marker GWIS, focusing only on SNPs with $p \leq 0.01$ for the GWAS analysis. Age, sex, and diagnosis were used as covariates in both analyses. Corrected p values using the Bonferroni method were reported. The GWAS analysis revealed significant hits within or proximal to *APOE*, *APOC1*, and *TOMM40* genes, which were previously implicated in AD. The GWIS analysis yielded 8 novel SNP-SNP interaction findings that warrant replication and further investigation.

1. Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disorder characterized by a progressive decline in memory and cognition. The pathologic cascade in AD involves two primary hallmarks: amyloid-β (Aβ) plaques and neurofibrillary tangles [1]. Genetics plays an important role in late-onset Alzheimer's disease (LOAD), but missing heritability remains to be found according to current approximations [2]. The last several decades of research yielded only one genetic risk factor of large effect for LOAD: Apolipoprotein E (*APOE*) with 2 copies of the $\varepsilon 4$ allele confers approximately 6- to 30-fold risk for the disease [3]. Some recent genome-wide association studies (GWAS) have identified several additional AD susceptibility genes, including *BIN1*, *CLU*, *ABCA7*, *CR1*, *PICALM*, *MS4A6A*, *MS4A4E*, *CD33*, *CD2AP*, and *EPHA1* [4–9]. However, these genetic factors have relatively low effect sizes (odds ratios of 0.87–1.23) and cumulatively account for

approximately 35% of population-attributable risk [8]. More recently, a large scale GWAS meta-analysis identified 11 new susceptibility loci with also small effect sizes [10].

Traditional GWAS analyses used discrete disease status as the phenotypic trait of interest despite the fact that LOAD is a clinically heterogeneous disorder. Recently, researchers started to explore intermediate quantitative traits (QTs), such as clinical or cognitive features, biochemical assays, or neuroimaging biomarkers, in genetic association testing. This may have the potential to address the issue of clinical heterogeneity in LOAD. These QTs are often measured as continuous variables and thus exhibit a higher genetic signal-to-noise ratio. Further, most intermediate QTs are more proximal to their genetic bases than disease status. Thus, the incorporation of intermediate QTs can potentially increase statistical power to detect disease-related genetic associations [11, 12]. An ancillary benefit of using QTs is that they can serve as effective biomarkers for monitoring disease progress or treatment response in clinical practice or drug trials.

Over the past 10–15 years, studies have identified robust and predictive biomarkers for AD including levels of tau and amyloid-β peptides in cerebrospinal fluid (CSF), selective measures of brain atrophy using magnetic resonance imaging (MRI), and imaging of glucose hypometabolism and amyloid using positron emission tomography (PET) [13]. PET imaging can be used to quantify levels of amyloid in the brain by utilizing a radiotracer such as florbetapir (^{18}F-AV-45 or AV-45) or/and Pittsburgh compound-B (PiB, N-methyl-[^{11}C]^2-(40-methylaminophenyl)-6-hydroxybenzothiazole). These amyloid measures have been studied as biomarkers for classifying AD [14–17]. All these multimodal biomarkers can potentially be served as AD relevant QTs and have been examined in many existing quantitative genetics studies of LOAD [18].

In addition, most genetic association studies compute statistics at the single marker level and ignore the possible underlying biological interactions that contribute to the development of disease [19] and could be a possible source for "missing heritability." Given the quadratically growing search space of two-way interactions, we are facing major computational and statistical challenges. To address this issue, one approach is to effectively explore epistatic interactions in genome-wide data by using a priori statistical and/or biological evidence to generate a reduced set of genetic markers for interaction testing. Using this strategy, previous interaction studies in LOAD (e.g., [20–24]) implicated interactions between *CR1* and *APOE* using quantified Aβ PET as the outcome variable [24] and between cholesterol trafficking genes [21, 22] and tau phosphorylation genes [20] in case-control analyses. These studies demonstrated that the important information could be garnered from investigating genetic interactions in complex diseases like LOAD.

With these observations, in the present work, we conducted a quantitative genetics study of an AD-associated amyloid imaging phenotype and examined both single marker main effects and two-marker interaction effects at the genome-wide level. Specifically, we investigated the main and interaction effects of genome-wide markers on cingulate amyloid-beta (Aβ) load in an effort to better understand the genetic etiology of cingulate cortical Aβ deposition (a LOAD biomarker).

2. Materials and Methods

Data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and nonprofit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, aged 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see http://www.adni-info.org/.

We applied for and were granted permission to use data from the ADNI cohort (http://www.adni-info.org/) to conduct genetic association and interaction analyses.

2.1. Subjects and Data. For the present work, analyses were restricted to subjects with both genotyping data and AV-45 PET data available. The study sample ($N = 602$) included 190 healthy control (HC), 215 early MCI (EMCI), 152 late MCI (LMCI), and 45 AD subjects. Table 1 shows selected demographic and clinical characteristics of these participants at the time of the baseline AV-45 PET scan.

2.2. Genotyping Data and Quality Control. The genotyping data of the participants were collected using either the Illumina 2.5 M array (a byproduct of the ADNI whole genome sequencing sample) or the Illumina OmniQuad array [18, 25, 26]. For the present analyses, we included single nucleotide

polymorphism (SNP) markers that were present on both arrays.

Quality control (QC) was performed using the PLINK software (version 1.07) [27]. SNPs not meeting any of the following criteria were excluded from further analyses: (1) call rate per SNP ≥95%; (2) minor allele frequency ≥ 5% (n = 117, 175 SNPs were excluded based on criteria 1 and 2); and (3) Hardy-Weinberg equilibrium test of $p \geq 10^{-6}$ (n = 997 SNPs were excluded) using control subjects only. Participants were excluded from the analysis if any of the following criteria were not satisfied: (1) call rate per participant ≥ 90% (3 participants were excluded); (2) sex check (1 participant was excluded); and (3) identity check for related pairs (3 sibling pairs were identified with PI_HAT >0.5; one participant of each pair was randomly selected and excluded from the study).

Population stratification analysis was performed using EIGENSTRAT [28] and confirmed using STRUCTURE [29]. It yielded 47 study participants who did not cluster with the remaining subjects and with the CEU HapMap samples who are primarily of European ancestry (non-Hispanic Caucasians). These 47 participants were excluded from the analysis. After QC, 582,718 SNPs and 602 samples remained available for genetic association and interaction analyses.

2.3. Quantitative Traits. A previous AV-45 PET study [30] showed that both AD and amnestic MCI subjects had higher standardized uptake value ratio (SUVR) in global cortical, precuneus, frontal, occipital, and posterior cingulate areas. We focused this study in one of these regions, which is cingulate. UC Berkeley extracted baseline SUVR mean measure from the cingulate cortical region (version 2014.7.30) that was downloaded from the ADNI database (http://adni.loni.usc.edu/) for 987 ADNI-GO/2 participants. We also downloaded the cerebellum SUVR measure and used it to normalize the cingulate SUVR measure. The normalized SUVR was used as the quantitative trait (QT) in our analyses. After excluding 383 participants due to the lack of genotyping data, 602 individuals remained in the further analysis.

In addition, amyloid-β 1-42 peptide (Aβ-42), total tau (t-tau), and tau phosphorylated at the threonine 181 (p-tau181p), measured in CSF samples, are potential diagnostic biomarkers for AD [31–33]. Among the 602 individuals, 504 have both AV-45 data and CSF data. Following a previous GWAS study on CSF biomarkers [34], QC was performed on the CSF data to reduce the potential influence of extreme outliers on statistical results. Mean and standard deviation (SD) of Aß1-42 and 2 ratios (t-tau/Aß1-42 and p-tau181p/Aß1-42) were calculated, blind to diagnostic information. Subjects who had at least one value greater or smaller than 4 SDs from the mean value of each of 3 CSF variables were regarded as extreme outliers and removed from the analysis. This step removed 5 additional participants, resulting in 499 valid CSF samples.

2.4. Genetic Association Studies: Main Effects and Interaction Effects. For GWAS examining the main effects, linear regression was performed using PLINK to determine the association of each SNP to the AV-45 measure. An additive genetic model was tested with covariates including age, gender, and diagnosis (through four binary dummy variables indicating HC, EMCI, LMCI, or AD). Manhattan plots and Q-Q plots were generated using Haploview (http://www.broad.mit.edu/mpg/haploview/) and R (http://www.r-project.org/), respectively.

For GWIS examining the interaction effects, the INTERSNP software [35] was applied to the genotyping data and phenotypic AV-45 measure. First, a single marker p value for the main effect was computed for each SNP. Top 10,000 SNPs with the smallest p values were selected and included in the subsequent interaction analysis. An explicit test for additive interaction (the full model including both additive and dominance effects plus interaction term versus reduced model that does not contain interaction terms) was performed on all possible SNP pairs among the top 10,000 SNPs, using two-marker analysis. The computation was conducted in a linear regression framework. We examined the association between SNP-SNP interactions and the AV-45 measure while controlling for relevant covariates at the baseline scan, including age, sex, and clinical diagnosis. This resulted in a total of approximately 50 million unique SNP pairs to be tested from the ADNI dataset. Interactions were considered significant if their Bonferroni corrected p value < 0.05.

2.5. Post Hoc Analysis. For identified significant interactions, we applied hierarchical linear regression using IBM SPSS 20 to estimate the amount of variance (R^2) in the AV-45 measure accounted for by these interaction terms. We first included the same set of covariates (age, gender, and diagnosis) in the linear model. After that, we included *APOE* status, the closest SNP to the *BCHE* SNP identified in a prior amyloid GWAS study [36], and the two SNP main effects from the identified SNP pair. Finally, we included the SNP-SNP interaction term to calculate additional variance explained by the interaction term. The difference in R^2 for the significant models was calculated in SPSS as $\Delta R^2 = R^2$ (full model with interaction term) $-R^2$ (reduced model without interaction term). Significant effects were plotted in SPSS as well.

In addition, based on the identified interactions associated with AV-45, we further evaluated their main and interaction effects on the CSF levels related to amyloid, including Aβ1-42, t-tau181p/Aβ1-42, and p-tau/Aβ1-42. These three CSF measures were used as the QTs in 3 separate genetic analyses, following the same method and steps for analyzing AV-45 phenotype as described above.

3. Results and Discussion

3.1. GWAS Results. Table 1 shows selected demographic and clinical characteristics of 602 ADNI participants analyzed in this study, where the EMCI group is slightly younger than the other groups. Figure 1 shows the Q-Q plot, indicating no evidence of spurious inflation. Figure 2 shows the Manhattan plot. As expected, significant associations were identified between loci on chromosome 19 and the AV-45 measure. The top association is from rs4420638 ($P = 5.11 \times 10^{-21}$), which codes for the *APOC1* [37]. A few other SNPs within the

Table 1: Selected demographic and clinical characteristics of participants at the time of AV-45 PET scan.

	HC (N = 190)	EMCI (N = 215)	LMCI (N = 152)	AD (N = 45)
Age (years)	74.51 (5.74)	71.43 (7.28)	73.03 (7.49)	74.87 (9.05)
Women	94 (49%)	95 (44%)	62 (41%)	17 (38%)
Education (years)	16.53 (2.64)	15.95 (2.66)	16.32 (2.90)	15.67 (2.70)
APOE e4 allele present	54 (28%)	87 (40%)	78 (51%)	33 (73%)
CDR-SOB	0.03 (0.13)	1.22 (0.72)	1.73 (0.94)	4.36 (1.64)
Mini mental status examination	29.07 (1.20)	28.39 (1.46)	27.25 (1.77)	22.93 (2.08)
Logical memory immediate recall (WMS-R)	14.46 (3.08)	10.96 (2.77)	7.32 (3.06)	4.40 (2.52)
Logical memory delayed recall (WMS-R)	13.55 (3.27)	8.90 (1.72)	4.22 (2.75)	2.02 (2.17)
Normalized SUVR of cingulate amyloid burden	1.211 (0.21)	1.273 (0.23)	1.274 (0.27)	1.48 (0.24)

AD: Alzheimer's disease; CDR–SOB: clinical dementia rating-sum of boxes; EMCI: early mild cognitive impairment; HC: healthy control; LMCI: late mild cognitive impairment; PET: positron emission tomography; WMS-R: Wechsler Memory Scale-Revised. Data are shown in the format of "number (%)" or "mean (SD)."

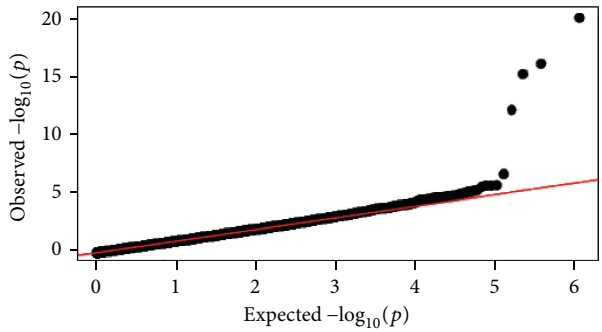


Figure 1: Quantile-quantile (Q-Q) plot of the observed $-\log_{10}p$ values from the GWAS of cingulate cortical Aß load versus those expected under the null hypothesis.

APOE region, including adjacent *APOC1* and *TOMM40*, were significantly associated with the AV-45 level in cingulate.

3.2. SNP-SNP Interaction Results. The INTERSNP model we tested included age, sex, and diagnosis as covariates. Eight SNP pairs showed significant interaction effects on the cingulate AV45 measure (corrected p value < 0.05) (Table 2): rs2194938 (*CLSTN2*)-rs7644138 (*FHIT*), rs7916162 (*TACC2*)-rs2326536 (*PRNP**), rs2295873 (*TACC2*)-rs7794838 (*IGFBP3**), rs2295874 (*TACC2*)-rs2326536 (*PRNP**), rs13056151 (*BCR*)-rs17594541 (*MAGI2*), rs13426621 (LOC388942)-rs7037332 (*TYRP1**), rs16936424 (LOC387761)-rs10504164 (*N/A*), and rs16939265 (*HNF4G**)-rs6854047 (*RWDD4**).

3.3. Post Hoc Analysis. Table 2 also shows the results of post hoc analysis on cingulate amyloid deposition. Age, gender, and diagnosis were first included in the model and accounted for 11% of variance in the amyloid QT. *APOE* status was then accounted for an additional 16.1% of variance, followed by the closest SNP to the *BCHE* SNP identified in [36] accounted for an additional 1.8% of variance. For each interaction, we ran a hierarchical linear regression model. We first added in the genetic main effects and then the genetic interaction term to determine the variance associated with the interaction term alone. For rs2194938 (*CLSTN2*)-rs7644138 (*FHIT*), the SNP main effects accounted for 3.4% of variance, and the interaction term accounted for 4.9% of variance (8.3% combined). For rs7916162 (*TACC2*)-rs2326536 (*PRNP**), the main effects accounted for 2% of variance, and the interaction accounted for 4.9% of variance (6.9% combined). For rs2295873 (*TACC2*)-rs7794838 (*IGFBP3**), the main effects accounted for 3.7% of variance, and the interaction term accounted for 4.1% of variance (7.8% combined). For rs2295874 (*TACC2*)-rs2326536 (*PRNP**), the SNP main effects accounted for 3.7% of variance, and the interaction term accounted for 4.1% of variance (7.8% combined). For rs13056151 (*BCR*)-rs17594541 (*MAGI2*), the main effects accounted for 3.5% of variance, and the interaction term accounted for 2.6% of variance (6.1% combined). For rs13426621 (LOC388942)-rs7037332 (*TYRP1**), the main effects accounted for 4.2% of variance, and the interaction accounted for 2.3% of variance (6.5% combined). For rs16936424 (LOC387761)-rs10504164 (*N/A*), the main effects accounted for 3.7% of variance, and the interaction term accounted for 1.7% of variance (5.4% combined). For rs16939265 (HNF4G*)-rs6854047 (RWDD4*), the main effects accounted for 2.7% of variance, and the interaction term accounted for 1.3% of variance (4.0% combined).

Using a slightly reduced sample (N = 499) with CSF biomarker data available, all 8 identified interactions remained statistically significant when performing hierarchical linear regression using the CSF phenotypes (one baseline measure: Aβ, two ratios: t-Tau/Aβ and p-Tau/Aβ) instead of the AV-45 measure as outlined earlier (Table 3). We also repeated the same AV-45 analysis on the reduced sample and achieved a very similar result (Table 4).

3.4. Discussion. In this study, we performed both GWAS and GWIS analyses of the cingulate AV-45 florbetapir PET measure, using a sample of 602 subjects from the ADNI database. To our knowledge, this is the first genome-wide study on examining SNP-SNP interaction effects on cingulate amyloid deposition in a substantially large sample. In the single marker analysis, as expected, SNPs in *APOE*, *APOC1*, and *TOMM40* genes (Figure 2) exhibited genome-wide

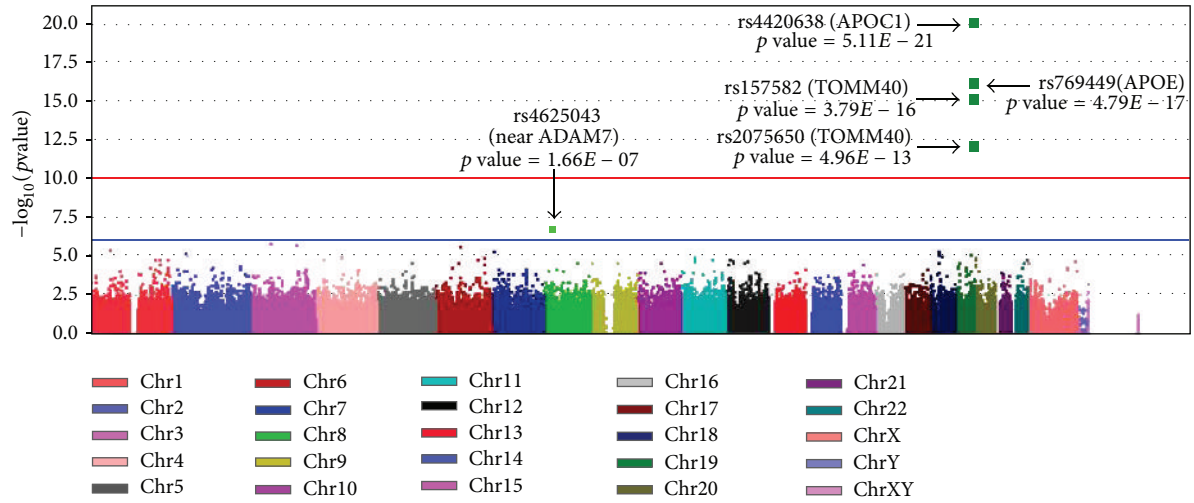


FIGURE 2: Manhattan plot of the observed $-\log_{10} p$ values from the GWAS of cingulate cortical Aβ load. More than 580,000 SNPs were tested for association with cingulate cortical Aβ burden under an additive model, with age, gender, and diagnosis as covariates. Genome-wide significant associations (exceeding the threshold represented by the red line and determined by Bonferroni correction) were identified on chromosome 19 within the *APOE* and its neighboring regions.

significant associations to the cingulate cortical Aß level. Two-marker interaction analyses revealed 8 SNP pairs, which had significant genetic interactions (corrected $p \leq 0.05$) with cingulate amyloid burden. The risk variants at these pairs had low main effects but explained a relatively high-level variance of the amyloid deposition in cingulate (Table 2).

In addition, missing heritability can partially be explained by the interaction effects that are not examined in traditional GWAS analyses. Genetic risk underlying diagnosis of LOAD is considered to be manifested from multiple genes which interact with each other. We have performed a post hoc analysis investigating the effects of the identified SNP-SNP interactions LOAD related quantitative phenotypes including amyloid deposition and CSF biomarkers (Aβ, t-tau/Aβ, p-tau/Aβ). Given amyloid and tau phosphorylation as major AD hallmarks, it is not surprising to observe the genetic interaction effects on both the amyloid load and relevant CSF biomarkers (Tables 2–4). Our results suggest that significant SNP-SNP interactions could exist between SNPs with low and insignificant main effects, and these interactions could be associated with altered amyloid burden and explain high-level risk in AD.

In line with our hypothesis, we identified multiple significant genetic interactions associated with cingulate amyloid deposition. Several genes found in this study have already been implicated in AD, thus lending confidence to the analytic procedure and results. These genes include *PRNP* [38, 39], *IGFBP3* [40, 41], and *MAGI2* [42, 43]. For example, Guerreiro et al. reported a nonsense mutation in *PRNP* associated with clinical Alzheimer's disease [38]. Ikonen et al. showed that interaction between the Alzheimer's survival peptide humanin and insulin-like growth factor-binding protein 3 (*IGFBP3*) regulates cell survival and apoptosis [40]. Potkin et al. identified an *MAGI2* SNP associated with hippocampal atrophy using the ADNI data [42]. Perhaps more importantly, this study also identified a number of SNPs that had not yet been associated with AD in conventional GWAS studies. Thus, this study exposes several potential candidate genes that could be explored in future replication samples.

This study had several methodological and technical advantages over other imaging genetics studies in addition to the above interesting findings. (1) To our knowledge this is the first genome-wide study to explore how SNP-SNP interactions influence cingulate amyloid burden, measured using florbetapir PET scan information. (2) Using continuous quantitative traits as phenotypes confers higher statistical power than using conventional clinical status. (3) The sample in this study included HC, EMCI, LMCI, and AD, thus providing a continuous and wide spectrum of the disease progression in the dataset. (4) Our approach embraced, rather than ignored, the confounding factors including age, sex, diagnosis, and previously identified risk genes *APOE* and *BCHE* and provided more accurate estimate of the interaction effects on amyloid burden. (5) CSF data were used in this study to cross-check the identified interactions, which had the potential to serve as an indirect validation strategy or provide complemental information.

Our study has several limitations. (1) We used single marker main effect value to select SNPs for interaction analysis, which could miss significant interactions between SNPs with insignificant main effects. (2) The small cell size in the interaction analyses might introduce false positives. (3) Our approach is mostly data-driven, without utilizing any existing biological knowledge (e.g., pathways, networks, and other functional annotation data), which may reduce the statistical power and result interpretability.

Table 2: Results of sample ($N = 602$): eight SNP-SNP interactions associated with cingulate amyloid burden. The Bonferroni corrected p values (<0.05) and R^2 of the SNP-SNP interaction term are shown in bold.

Number	SNP1 × SNP2	Gene	CHR	Main effect	Interaction		*R* square				
				p value	p value	Corrected p value	Age + Sex + Dx[a]	*APOE*[b]	*BCHE*[c]	SNP1 + SNP2[d]	SNP1 ∗ SNP2[e]
1	rs2194938 ×	CLSTN2	3	0.000481499	$5.24E-10$	**0.026**	0.110	0.161	0.018	0.034	**0.049**
	rs7644138	FHIT	3	0.000993424							
2	rs7916162 ×	TACC2	10	0.00897357	$7.81E-10$	**0.038**	0.110	0.161	0.018	0.020	**0.049**
	rs2326536	PRNP*	20	0.00850742							
3	rs2295873 ×	TACC2	10	0.000291361	$7.01E-10$	**0.035**	0.110	0.161	0.018	0.037	**0.041**
	rs7794838	IGFBP3*	7	0.00973379							
4	rs2295874 ×	TACC2	10	0.0016572	$8.62E-10$	**0.042**	0.110	0.161	0.018	0.024	**0.039**
	rs2326536	PRNP*	20	0.00850742							
5	rs13056151 ×	BCR	22	0.0015002	$9.86E-10$	**0.048**	0.110	0.161	0.018	0.035	**0.026**
	rs17594541	MAGI2	7	0.00360174							
6	rs13426621 ×	LOC388942	2	$6.72E-06$	$3.44E-10$	**0.017**	0.110	0.161	0.018	0.042	**0.023**
	rs7037332	TYRP1*	9	0.00386214							
7	rs16936424 ×	LOC387761	11	$1.24E-05$	$9.22E-10$	**0.045**	0.110	0.161	0.018	0.037	**0.017**
	rs10504164	NA	8	0.00166034							
8	rs16939265 ×	HNF4G*	8	0.000407669	$6.95E-10$	**0.034**	0.110	0.161	0.018	0.027	**0.013**
	rs6854047	RWDD4*	4	0.000464343							

[a] Age + Sex + Dx: percent of variance in cingulate amyloid burden explained by age, gender, and diagnosis.

[b] *APOE*: percent of additional variance in cingulate amyloid burden explained by the APOE genotype after accounting for age, gender, and diagnosis.

[c] *BCHE*: percent of additional variance in cingulate amyloid burden explained by the *BCHE* SNP after accounting for age, gender, diagnosis, and *APOE* genotype.

[d] SNP1 + SNP2: percent of additional variance in cingulate amyloid burden explained by the combined main effect of SNP1 and SNP2 after accounting for age, gender, diagnosis, *APOE* genotype, and the *BCHE* SNP.

[e] SNP1 ∗ SNP2: percent of additional variance in cingulate amyloid burden explained by the interaction effect of SNP1 and SNP2 after accounting for age, gender, diagnosis, *APOE* genotype, the *BCHE* SNP, SNP1, and SNP2.

* Nearest gene proximal to the SNP.

Table 3: Results of sample (N = 499): eight SNP-SNP interaction associations with three CSF biomarkers.

Number	SNP1 × SNP2	Aβ (R square)					t-Tau/Aβ (R square)					p-Tau/Aβ (R square)				
		Age + Sex + Dx	APOE	BCHE	SNP1 + SNP2	SNP1 * SNP2	Age + Sex + Dx	APOE	BCHE	SNP1 + SNP2	SNP1 * SNP2	Age + Sex + Dx	APOE	BCHE	SNP1 + SNP2	SNP1 ∗ SNP2
1	rs2194938 × rs7644138	0.108	0.187	0.012	0.021	0.008	0.132	0.153	0.022	0.016	0.014	0.134	0.129	0.024	0.007	0.011
2	rs7916162 × rs2326536	0.108	0.187	0.012	0.003	0.019	0.132	0.153	0.022	0.004	0.011	0.134	0.129	0.024	0.005	0.014
3	rs2295873 × rs7794838	0.108	0.187	0.012	0.014	0.012	0.132	0.153	0.022	0.017	0.011	0.134	0.129	0.024	0.016	0.007
4	rs2295874 × rs2326536	0.108	0.187	0.012	0.003	0.018	0.132	0.153	0.022	0.002	0.012	0.134	0.129	0.024	0.003	0.015
5	rs13056151 × rs17594541	0.108	0.187	0.012	0.005	0.005	0.132	0.153	0.022	0.005	0.006	0.134	0.129	0.024	0.005	0.004
6	rs13426621 × rs7037332	0.108	0.187	0.012	0.002	0.002	0.132	0.153	0.022	0.006	0.012	0.134	0.129	0.024	0.004	0.006
7	rs16936424 × rs10504164	0.108	0.187	0.012	0.018	0.004	0.132	0.153	0.022	0.014	0.008	0.134	0.129	0.024	0.009	0.003
8	rs16939265 × rs6854047	0.108	0.187	0.012	0.012	0.007	0.132	0.153	0.022	0.003	0.005	0.134	0.129	0.024	0.002	0.003

Table 4: Results of sample ($N = 499$): eight SNP-SNP interactions associated with cingulate amyloid burden.

Number	SNP1 × SNP2	Gene	CHR	Main effect	Interaction		*R* square				
				p value	*p* value	corrected *p* value	Age + Sex + Dx	APOE	BCHE	SNP1 + SNP2	SNP1 ∗ SNP2
1	rs2194938 ×	CLSTN2	3	0.000481499	$5.24E-10$	**0.026**	0.127	0.133	0.019	0.047	**0.053**
	rs7644138	FHIT	3	0.000993424							
2	rs7916162 ×	TACC2	10	0.00897357	$7.81E-10$	**0.038**	0.127	0.133	0.019	0.027	**0.055**
	rs2326536	PRNP*	20	0.00850742							
3	rs2295873 ×	TACC2	10	0.000291361	$7.01E-10$	**0.035**	0.127	0.133	0.019	0.059	**0.059**
	rs7794838	IGFBP3*	7	0.00973379							
4	rs2295874 ×	TACC2	10	0.0016572	$8.62E-10$	**0.042**	0.127	0.133	0.019	0.031	**0.046**
	rs2326536	PRNP*	20	0.00850742							
5	rs13056151 ×	BCR	22	0.0015002	$9.86E-10$	**0.048**	0.127	0.133	0.019	0.034	**0.037**
	rsl7594541	MAGI2	7	0.00360174							
6	rs13426621 ×	LOC388942	2	$6.72E-06$	$3.44E-10$	**0.017**	0.127	0.133	0.019	0.049	**0.039**
	rs7037332	TYRP1*	9	0.00386214							
7	rs16936424 ×	LOC387761	11	$1.24E-05$	$9.22E-10$	**0.045**	0.127	0.133	0.019	0.051	**0.023**
	rs10504164	NA	8	0.0166034							
8	rs16939265 ×	HNF4G*	8	0.000407669	$6.95E-10$	**0.034**	0.127	0.133	0.019	0.040	**0.026**
	rs6854047	RWDD4*	4	0.000464343							

*Nearest gene proximal to the SNP.

4. Conclusions

We performed GWAS and GWIS using amyloid imaging as the quantitative phenotype and investigated the genetic interaction effects on cingulate amyloid-beta (Aβ) load. The single marker analyses revealed significant hits within or proximal to *APOE*, *APOC1,* and *TOMM40* genes, which were previously implicated in AD. The interaction analyses yielded a few novel interaction findings associated with cingulate amyloid burden, such as those between *CLSTN2* and *FHIT*, between *TACC2* and *PRNP*, between *TACC2* and *IGFBP3*, and between *BCR* and *MAGI2*. Each of these SNP pairs demonstrated significant interaction effects while their individual main effects were not prominent. This suggests that searching for interaction effects may help solve the problem of missing heritability to some extent. Future studies should attempt to replicate these results in independent datasets with neuroimaging and genetic data, as they become available. Additional pathway analysis and gene sets enrichment analysis could be performed to help understand the genetic interactions between SNPs on amyloid imaging phenotypes and potentially provide critical functional evidence in support of the statistical association findings.

Conflict of Interests

The authors have no actual or potential conflict of interests including any financial, personal, or other relationships with other people or organizations that could inappropriately influence (bias) their work.

Acknowledgments

This paper is supported in part by grants from National Natural Science Foundation of China (61471139, 61403092, and 61071174), National Key Scientific Instrument and Equipment Development Projects of China (2012YQ04014001 and 2012YQ04014010), Fundamental Research Funds for the Central Universities (HEUCFT1102, HEUCFT1302, and HEUCFX41303), and the HEU 111 Project at Harbin Engineering University and by NIH R01 LM011360, U01 AG024904, RC2 AG036535, R01 AG19771, P30 AG10133, UL1 TR001108, NSF IIS-1117335, DOD W81XWH-14-2-0151, and NCAA 14132004 at IU. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense Award no. W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org/). The grantee organization is the Northern Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data used in preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this paper. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

[1] D. L. Price and S. S. Sisodia, "Mutant genes in familial Alzheimer's disease and transgenic models," *Annual Review of Neuroscience*, vol. 21, no. 1, pp. 479–505, 1998.

[2] L. Bertram, C. M. Lill, and R. E. Tanzi, "The genetics of alzheimer disease: back to the future," *Neuron*, vol. 68, no. 2, pp. 270–281, 2010.

[3] H. Akiyama, K. Ikeda, H. Kondo, M. Kato, and P. L. McGeer, "Microglia express the type 2 plasminogen activator inhibitor in the brain of control subjects and patients with Alzheimer's disease," *Neuroscience Letters*, vol. 164, no. 1-2, pp. 233–235, 1993.

[4] D. Harold, R. Abraham, P. Hollingworth et al., "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease," *Nature Genetics*, vol. 41, pp. 1088–1093, 2009.

[5] O. Belbin, M. M. Carrasquillo, M. Crump et al., "Investigation of 15 of the top candidate genes for late-onset Alzheimer's disease," *Human Genetics*, vol. 129, no. 3, pp. 273–282, 2011.

[6] M. M. Carrasquillo, T. A. Hunter, L. Ma et al., "Replication of BIN1 association with Alzheimer's disease and evaluation of genetic interactions," *Journal of Alzheimer's Disease*, vol. 24, no. 4, pp. 751–758, 2011.

[7] P. Hollingworth, D. Harold, R. Sims et al., "Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease," *Nature Genetics*, vol. 43, no. 5, pp. 429–436, 2011.

[8] A. C. Naj, G. Jun, G. W. Beecham et al., "Common variants at *MS4A4/MS4A6E*, *CD2AP*, *CD33* and *EPHA1* are associated with late-onset Alzheimer's disease," *Nature Genetics*, vol. 43, no. 5, pp. 436–441, 2011.

[9] H. Shi, O. Belbin, C. Medway et al., "Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS)," *Neurobiology of Aging*, vol. 33, no. 8, pp. e5–e18, 2012.

[10] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold et al., "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Nature Genetics*, vol. 45, no. 12, pp. 1452–1458, 2013.

[11] N. F. Bence, R. M. Sampat, and R. R. Kopito, "Impairment of the ubiquitin-proteasome system by protein aggregation," *Science*, vol. 292, no. 5521, pp. 1552–1555, 2001.

[12] S. G. Potkin, J. A. Turner, G. Guffanti et al., "Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations," *Cognitive Neuropsychiatry*, vol. 14, no. 4-5, pp. 391–418, 2009.

[13] B. Dubois, H. H. Feldman, C. Jacova et al., "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria," *The Lancet Neurology*, vol. 6, no. 8, pp. 734–746, 2007.

[14] M. D. Ikonomovic, W. E. Klunk, E. E. Abrahamson et al., "Post-mortem correlates of *in vivo* PiB-PET amyloid imaging in a typical case of Alzheimer's disease," *Brain*, vol. 131, no. 6, pp. 1630–1645, 2008.

[15] M. S. Albert, S. T. DeKosky, D. Dickson et al., "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's and Dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[16] C. M. Clark, J. A. Schneider, B. J. Bedell et al., "Use of florbetapir-PET for imaging beta-amyloid pathology," *The Journal of the American Medical Association*, vol. 305, no. 3, pp. 275–283, 2011.

[17] R. A. Sperling, P. S. Aisen, L. A. Beckett et al., "Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[18] L. Shen, P. M. Thompson, S. G. Potkin et al., "Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers," *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 183–207, 2014.

[19] S. A. Meda, M. E. I. Koran, J. R. Pryweller, J. N. Vega, and T. A. Thornton-Wells, "Genetic interactions associated with 12-month atrophy in hippocampus and entorhinal cortex in Alzheimer's Disease Neuroimaging Initiative," *Neurobiology of Aging*, vol. 34, no. 5, pp. 1518.e9–1518.e18, 2013.

[20] I. Mateo, J. L. Vázquez-Higuera, P. Sánchez-Juan et al., "Epistasis between tau phosphorylation regulating genes (CDK5R1 and GSK-3β) and Alzheimer's disease risk," *Acta Neurologica Scandinavica*, vol. 120, no. 2, pp. 130–133, 2009.

[21] E. Rodríguez-Rodríguez, I. Mateo, J. Infante et al., "Interaction between HMGCR and ABCA1 cholesterol-related genes modulates Alzheimer's disease risk," *Brain Research*, vol. 1280, pp. 166–171, 2009.

[22] E. Rodriguez-Rodriguez, J. L. Vázquez-Higuera, P. Sánchez-Juan et al., "Epistasis between intracellular cholesterol trafficking-related genes (NPC1 and ABCA1) and Alzheimer's disease risk," *Journal of Alzheimer's Disease*, vol. 21, no. 2, pp. 619–625, 2010.

[23] M. E. I. Koran, T. J. Hohman, and T. A. Thornton-Wells, "Genetic interactions found between calcium channel genes modulate amyloid load measured by positron emission tomography," *Human Genetics*, vol. 133, no. 1, pp. 85–93, 2014.

[24] M. Thambisetty, Y. An, M. Nalls et al., "Effect of complement CR1 on brain amyloid burden during aging and its modification by APOE genotype," *Biological Psychiatry*, vol. 73, no. 5, pp. 422–428, 2013.

[25] A. J. Saykin, L. Shen, T. M. Foroud et al., "Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans," *Alzheimer's & Dementia*, vol. 6, no. 3, pp. 265–273, 2010.

[26] L. Shen, S. Kim, S. L. Risacher et al., "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort," *NeuroImage*, vol. 53, no. 3, pp. 1051–1063, 2010.

[27] S. Purcell, B. Neale, K. Todd-Brown et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[28] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[29] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[30] K.-L. Huang, K.-J. Lin, I.-T. Hsiao et al., "Regional amyloid deposition in amnestic mild cognitive impairment and Alzheimer's disease evaluated by [^{18}F]AV-45 positron emission tomography in Chinese population," *PLoS ONE*, vol. 8, no. 3, Article ID e58974, 2013.

[31] W. J. Jagust, S. M. Landau, L. M. Shaw et al., "Relationships between biomarkers in aging and dementia," *Neurology*, vol. 73, no. 15, pp. 1193–1199, 2009.

[32] L. M. Shaw, H. Vanderstichele, M. Knapik-Czajka et al., "Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects," *Annals of Neurology*, vol. 65, no. 4, pp. 403–413, 2009.

[33] H. Hampel, K. Blennow, L. M. Shaw, Y. C. Hoessler, H. Zetterberg, and J. Q. Trojanowski, "Total and phosphorylated tau protein as biological markers of Alzheimer's disease," *Experimental Gerontology*, vol. 45, no. 1, pp. 30–40, 2010.

[34] S. Kim, S. Swaminathan, L. Shen et al., "Genome-wide association study of CSF biomarkers Aβ1-42, t-tau, and p-tau181p in the ADNI cohort," *Neurology*, vol. 76, no. 1, pp. 69–79, 2011.

[35] C. Herold, M. Steffens, F. F. Brockschmidt, M. P. Baur, and T. Becker, "INTERSNP: genome-wide interaction analysis guided by a priori information," *Bioinformatics*, vol. 25, no. 24, pp. 3275–3281, 2009.

[36] V. K. Ramanan, S. L. Risacher, K. Nho et al., "APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study," *Molecular Psychiatry*, vol. 19, no. 3, pp. 351–357, 2014.

[37] E. H. Corder, A. M. Saunders, W. J. Strittmatter et al., "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," *Science*, vol. 261, no. 5123, pp. 921–923, 1993.

[38] R. Guerreiro, J. Brás, A. Wojtas, R. Rademakers, J. Hardy, and N. Graff-Radford, "A nonsense mutation in PRNP associated with clinical Alzheimer's disease," *Neurobiology of Aging*, vol. 35, no. 11, pp. 2656.e13–2656.e16, 2014.

[39] C. Sassi, R. Guerreiro, R. Gibbs et al., "Investigating the role of rare coding variability in Mendelian dementia genes (*APP*, *PSEN1*, *PSEN2*, *GRN*, *MAPT*, and *PRNP*) in late-onset Alzheimer's disease," *Neurobiology of Aging*, vol. 35, no. 12, pp. 2881.e1–2881.e6, 2014.

[40] M. Ikonen, B. Liu, Y. Hashimoto et al., "Interaction between the Alzheimer's survival peptide humanin and insulin-like growth factor-binding protein 3 regulates cell survival and apoptosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 22, pp. 13042–13047, 2003.

[41] P. Kochunov, J. Charlesworth, A. Winkler et al., "Transcriptomics of cortical gray matter thickness decline during normal aging," *NeuroImage*, vol. 82, pp. 273–283, 2013.

[42] S. G. Potkin, G. Guffanti, A. Lakatos et al., "Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease," *PLoS ONE*, vol. 4, no. 8, Article ID e6501, 2009.

[43] M. W. Weiner, D. P. Veitch, P. S. Aisen et al., "The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception," *Alzheimer's and Dementia*, vol. 9, no. 5, pp. e111–e194, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 570568, 12 pages
http://dx.doi.org/10.1155/2015/570568

Research Article

How to Isolate a Plant's Hypomethylome in One Shot

Elisabeth Wischnitzki, Eva Maria Sehr, Karin Hansel-Hohl, Maria Berenyi, Kornel Burg, and Silvia Fluch

AIT Austrian Institute of Technology GmbH, Konrad-Lorenzstreet 24, 3430 Tulln, Austria

Correspondence should be addressed to Elisabeth Wischnitzki; elisabeth.wischnitzki@ait.ac.at

Received 5 December 2014; Revised 24 March 2015; Accepted 30 March 2015

Academic Editor: Weixing Feng

Genome assembly remains a challenge for large and/or complex plant genomes due to their abundant repetitive regions resulting in studies focusing on gene space instead of the whole genome. Thus, DNA enrichment strategies facilitate the assembly by increasing the coverage and simultaneously reducing the complexity of the whole genome. In this paper we provide an easy, fast, and cost-effective variant of MRE-seq to obtain a plant's hypomethylome by an optimized methyl filtration protocol followed by next generation sequencing. The method is demonstrated on three plant species with knowingly large and/or complex (polyploid) genomes: *Oryza sativa*, *Picea abies*, and *Crocus sativus*. The identified hypomethylomes show clear enrichment for genes and their flanking regions and clear reduction of transposable elements. Additionally, genomic sequences around genes are captured including regulatory elements in introns and up- and downstream flanks. High similarity of the results obtained by a *de novo* assembly approach with a reference based mapping in rice supports the applicability for studying and understanding the genomes of nonmodel organisms. Hence we show the high potential of MRE-seq in a wide range of scenarios for the direct analysis of methylation differences, for example, between ecotypes, individuals, within or across species harbouring large, and complex genomes.

1. Introduction

Chemical modifications of DNA and histones, known as epigenetic marks, regulate the access to the genetic information encoded in the DNA of eukaryotic cells. Thereby, epigenetic modifications can inheritably coordinate gene expression without changing the underlying DNA sequence. As such, epigenetic regulation is an additional layer in the genetic information of a cell influencing a plethora of biological processes [1, 2]. In plants, the most common mark of DNA methylation is 5-methylcytosine (5-mC) [3]. The cytosine can be methylated at CG, CHG, and CHH sites, where H represents nonguanine residues. Cytosine methylation is nonrandomly distributed in plants and is found primarily in repetitive regions of the genome that are enriched in transposable elements (TEs), centromeric repeats, or silent rDNA repeats. When DNA methylation occurs in promoter regions and within the gene space it is associated with differential gene expression [4, 5].

Based on whole genome DNA methylation analyses it is now widely accepted that methylation marks in plants fluctuate according to the cell, tissue, and organ in the vegetative and reproductive phases of a plant's life cycle [6, 7]. This epigenetic variation is of utmost importance not only during plant development but also in the response to environmental conditions. Most notably, cytosine methylation patterns acquired in response to abiotic or biotic stress are often inherited over one to several subsequent generations. Thereby, the epigenetic system reversibly stores information over time functioning as a "molecular memory." This transgenerational inheritance of DNA methylation can in some cases lead to novel epialleles and phenotypes within populations and thereby mediates phenotypic plasticity [8].

Thus, epigenetic profiling is an increasingly popular strategy for understanding the genetic and environmental interactions behind many biological processes. Therefore, robust, cost-effective, and scalable assays are needed for studying epigenetic variation in diverse contexts. Over

the past years numerous methods have been developed to study a plant's methylome (the methylated part of the genome) and hypomethylome (the nonmethylated part of the genome), whereby each method is accompanied by its strengths and limitations (reviewed in [9, 10]). Nowadays, sequencing-based methods especially present a unique opportunity to achieve comprehensive methylome or hypomethylome coverage.

The scientific goal to focus the sequencing efforts led to strategies to enrich either methylated or nonmethylated DNA regions. Immunoprecipitation followed by sequencing (MeDIP-seq) is used to obtain the methylated parts of genomes [11]. Due to the relatively low cost for acquiring genome-wide data, MeDIP-seq is very attractive and has recently been applied to complex plant genomes, such as poplar [12], maize [13], and rice [14]. One the contrary, to enrich the nonmethylated part of a genome (the hypomethylome), methylation-sensitive restriction enzymes have been used. Based on the fact that the gene body in plants is showing rather low methylation levels (hypomethylated) and that, in contrast, cytosine methylation is found predominantly in repetitive elements (e.g., transposable elements) [4], methylation-sensitive enzyme-based genome digests creating reduced representation library allow enriching gene related sequences [15, 16]. A widely applied variation of this methyl filtration (MF) approach is using the enzyme McrBC followed by cloning steps [17, 18]. The combination of MF with subsequent next generation sequencing (NGS) is termed MRE-seq (methylation-sensitive restriction enzyme-seq). This method has so far been predominantly applied in mammalian tissue for analysing methylation differences [19–21]. Although an enhanced MF method has been described in 2009 for plants [22], most of the recent studies in plants still study the hypomethylome through the McrBC-based MF [23–25], MSAP (methylation-sensitive amplified polymorphism [26, 27]), RLGS (Restriction Landmark Genome Scanning [28]), or methylation-sensitive Southern blotting [29].

Due to some limitations in MF techniques (reviewed by [9]), there is still potential to improve the MRE-seq in order to allow a wider application of the technique for the direct analysis of methylation differences between ecotypes and the role of epigenetics as a source of variation contributing to fitness and natural selection especially with regard to nonmodel organisms.

With the present study performed on the model organism rice (*Oryza sativa*) we demonstrate that with an improved MRE-seq method, the hypomethylome and thus the gene space of a plant can be easily accessed by the use of methylation-sensitive restriction enzymes followed by next generation sequencing. Using different bioinformatics approaches we show that performing *de novo* assembly with the MF sequences allows the reconstruction of a large proportion of the gene space including promoters without prior knowledge of the whole genome. Furthermore we confirm our results in small scale studies in the large genome of Norway spruce (*Picea abies*) and the triploid saffron crocus (*Crocus sativus*) genome.

Our method provides an easy tool for killing two birds with one stone: (1) the reduced representation library enriched for gene space can serve as cost-effective tool for analysing a plant's gene space depleted of repetitive elements comprising over 50–80% of the genome [30]; (2) with this representation of the hypomethylome, an easy comparative analysis of epigenetic variation among genotypes or tissues can be performed at an affordable price, even in a larger set of samples.

2. Material and Methods

2.1. Plant Material. Genomic DNA was prepared from leaves of the *Oryza sativa* ssp. *indica* variety SHZ-2A (seeds are kindly provided by R. Mauleon, IRRI International Rice Research Institute, Los Banos, Philippines), from a pool of stigmata of the *Crocus sativus* L. accession "LaMancha" (material kindly provided by O. Santana-Méridas, Servicios Periféricos de Agricultura, Centro Agrario de Albaladejito, Cuenca, Spain), and from needles of *Picea abies* (L.) H. Karst. (twigs kindly provided by S. Schüler, Department of Forest Genetics, Austrian Research Centre for Forests, Vienna, Austria) using DNeasy Plant mini kit (Qiagen) following the manufacturer's instructions.

2.2. Methyl Filtration with Size Selection through PCR

2.2.1. Enzyme Selection. In order to improve the MF enrichment towards a higher coverage of the gene space and to adjust the previously reported technique [22] towards NGS, the enrichment potency of five different methylation-sensitive enzymes (AciI, HpaII, and Bsh1236I sensitive to CpG methylation, and MspI and PspGI sensitive to CpH/WpG methylation) was evaluated in a first step using rice as the model of choice by following the steps described below.

2.2.2. Digestion and Ligation. Digestion of the genomic DNA and ligation of the adapters was performed simultaneously in a single reaction for each enzyme separately. 300 ng of genomic DNA and 4 μL 10 mM of the preannealed adaptors PmeI_CGWA (5′-GCACGACTGTTTAAA-3′) and PmeI_CGB (5′-CGTTTAAACAGTCGT-3′, 5′ phosphorylated) were mixed in 50 μL reaction volume supplemented either with AciI, Bsh1236I, HpaII, MspI, or PspGI (40 U, NEB) each in the corresponding NEB buffer. During the enzymatic digestion process, the cut DNA fragments were simultaneously ligated to the double stranded adaptors with 2 μL T4 ligase (Thermo Scientific) and 2 mM ATP being present in the same reaction mix. After overnight incubation at 37°C, the reaction was stopped by heat inactivation at 65°C for 20 minutes and diluted 1 : 1 with water. Samples were extracted with phenol-chloroform followed by chloroform before precipitation with EtOH. Samples were dissolved in 100 μL 0.5x NEB4 buffer.

2.2.3. Amplification of the Adaptor Ligated DNA. For Illumina sequencing, fragments were attained by PCR amplification of the restriction digested and adaptor ligated genomic DNA samples. 1 μL of digested and adaptor ligated DNA

and 6 μL of 10 μM amplification primer PmeI_CG17 (5′-CACGACTGTTTAAACGG-3′) were used in a 50 μL PCR reaction containing 2.5 U HotStart Polymerase (Qiagen), 1 μL 25 mM $MgCl_2$, and 1 μL 20 μM dNTPs. The PCR yielded 200–800 bp fragments under the following cycling conditions: 95°C for 15 minutes; 30 times 95°C/30 sec, 55°C/40 sec, 72°C/50 sec; 72°C for 5 minutes. The PCR reactions were precipitated in EtOH and DNA dissolved in 100 μL 5 mM Tris buffer (pH 8.0). Eight parallel reactions were performed for each restriction enzyme setup in order to collect sufficient amount of DNA for subsequent sequencing.

2.2.4. Removal of the Adaptor Sequences. To increase the length of the usable sequence information, the majority of the adaptor sequence was removed by PmeI digestion, the rare cutter site (GTTTAAAC) included in the adaptor sequence. 20 μg of the PCR amplifications were digested with PmeI (NEB) in NEB4 buffer and supplemented with 100 ng/μL BSA in two steps. First digestion was performed in a 200 μL reaction volume, containing 200 U PmeI enzyme on 37°C for 2 hours followed by a subsequent volume increase to 250 μL including additional 50 U PmeI and incubated for additional 2 hours. Finally the reaction was stopped at 65°C for 20 minutes.

2.3. Sequencing. The rice and Norway spruce fragments have been prepared as amplicon libraries and next generation sequencing was performed on Illumina's HiSeq 2000 using 100 bp paired end technology. The individual libraries (5 libraries of rice, each treated with one of the above mentioned enzymes, and 5 libraries of N. spruce, 4 treated with HpaII, and a whole genome snapshot library as control) were barcoded and sequenced together in a single lane. Library preparation and sequencing was done by GATC Biotech AG. The sequencing of the saffron crocus fragments was performed on an Illumina MiSeq machine (300 bp paired end reads). The library was prepared and barcoded using the TruSeq DNA PCR-Free LT Sample Preparation Kit has been quantified using the KAPA Library Quantification Kit on a standard qRT-PCR machine, and the quality has been checked on the Agilent Bioanalyzer using the Agilent High Sensitivity DNA kit. All kits have been applied according to the manufacturer's protocols. The sample was sequenced using the MiSeq Reagent Kit v3 according to manufacturer's protocols (Illumina Inc.) together with one other sample.

2.4. Sequencing Data Processing. All sequence reads were cleaned in order to guarantee high quality data by removing adaptor fragments, low quality regions (Q30), and short sequences (<50 bp; <100 bp for saffron crocus MiSeq data) from the datasets using *in house* developed Perl scripts. Then sequence reads were analysed for their origin from potential repetitive elements (REdat version 9.3 [31]), ribosomal data (*in house* reference database based on ribosomal data from NCBI and unpublished *in house* data), and chloroplast or mitochondrion DNA (*Oryza sativa* ssp. *japonica* release 7 [32], *Picea abies* release 1.0 [33]). Due to the lacking genome sequence of saffron crocus the rice genomic data was used as reference for this step. The TE-related reads were included in the further analysis to avoid an artificial bias against TEs.

All coverage calculations were performed by dividing the sum of base pairs of the respective dataset by the size of the studied sequences. The simulation to estimate the minimal coverage necessary to identify the hypomethylome of the whole genome was performed on the rice dataset by randomly selecting reads from the combined dataset with subsequent mapping to the genome sequence (*Oryza sativa* ssp. *japonica* release 7 [32]) using bowtie2 with default settings [34]. The identified regions were compared to the genomic area covered with the complete combined dataset using bedtools (version: 2.17.0; [35]) and the resulting overlap was calculated.

2.5. Additional Analysis of Genome Sequences. The separation of gene models into genes and TEs, the identification of the 1.000 bp up- and downstream flanking regions as well as exon and intron regions in the analysed genomes, is based on the annotation of genome release 7 for rice and genome release 1.0 for Norway spruce.

Frequencies of AciI and HpaII restriction sites in the genomic sequence of rice [32], *Arabidopsis* (TAIR10; [36]), poplar (JGI 2.0; [37]), grapevine (Genoscope_v1; [38]), Norway spruce (v1.0; [33]), maize (5b.60; [39]), sorghum (JGI 1.4; [40]), and *Brachypodium* (MIPS 1.2; [41]) were calculated using *in house* developed Perl scripts. The information of AciI and HpaII frequencies in *Homo sapiens* was taken from http://tools.neb.com/~posfai/TheoFrag/TheoreticalDigest.human.html.

Known regulatory elements from publicly available resources (JASPAR, Agris, AthaMap, Transfac, PLACE [42–46]) were filtered for degenerated sites and a minimal length of eight nucleotides to minimize the probability for the detection of nonfunctional patterns due to random occurrences within the sequences. The remaining elements were located in the rice genome and assigned to the identified regions using *in house* developed Perl scripts.

2.6. Reference Based and De Novo Assembly. The reference guided assembly was performed by assigning all high quality read sequences to the genome sequences of rice [32] or Norway spruce [33], respectively, using bowtie2 with default settings [34] resulting in regions representing the hypomethylated fragments. This was performed for each of the datasets separately. In order to guarantee that the regions used during downstream analyses do not represent false positives due to problems during the mapping, sequencing errors or technical problems during wet lab processes, only genomic regions were retained comprising at least five reads.

The *de novo* assembly for each enzyme and the combined dataset for rice and the dataset of saffron crocus was performed using Trinity [47] with a minimal contig length of 100 bp. The resulting contigs were evaluated by mapping the high quality reads used for the assembly to the assembled contig sequences using bowtie2 [34] and only contigs consisting of at least five reads were retained, similar to the reference based assembly. The contigs assembled for the rice datasets were compared to the rice reference genome

using blast (version: 2.2.21; *e*-value < $1e-20$; [48]) and genomic coordinates were assigned to each contig based on the best blast hit. For multiple occurrences with identical hit-statistics both entries were retained (10% of contigs). Not located contigs were subjected to a comparison to the NT database of NCBI using blast (version: 2.2.21; *e*-value < $1e-20$; [48]).

2.7. Comparative Sequence Analysis. Genomic coordinates of the reference based and *de novo* assembly were combined using bedtools (version: 2.17.0; [35]) and *in house* developed Perl scripts. Visualizations of the read location in the genome were created using the Integrated Genome Viewer (IGV 2.3; [49, 50]).

Additional methylation datasets for rice from the publications of He et al. [51], Yan et al. [14], and Li et al. [52] were derived from the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/). The retrieved raw read information was subjected to the previously described preprocessing procedures and only reads with quality scores of more than Q30 and minimal length of 20 bp were used for further analysis. The reads were mapped with the previous described procedure to the rice genome.

The *de novo* assembled contigs of the saffron crocus dataset were compared to the protein sequences or rice, maize (5b.60; [39]), and *Brachypodium* (MIPS 1.2; [41]) using blast (version: 2.2.21; *e*-value < $1e-10$; [48]).

3. Results and Discussion

3.1. Enzyme Selection. In a first analysis five different methylation-sensitive enzymes were analysed for their enrichment potency. AciI, HpaII, and Bsh1236I are sensitive to CpG methylation, while MspI and PspGI are sensitive to CpH/WpG methylation. Although the restriction sites of the enzymes differ (AciI (CCGC), HpaII (CCGG), Bsh1236I (CGCG), MspI (CCGG), and PspGI (CCWGG)), they are present in almost every gene and transposable element allowing a genome wide study (Figure 1). For all enzymes the resulting fragments were isolated and sequenced and the obtained reads were mapped to the genomic sequence of rice. Considering all reads the resulting hypomethylated regions identified about 90% of the annotated gene models of rice for all five enzymes. A clear depletion of transposable elements was observed for the three CpG methylation-sensitive enzymes AciI, Bsh1236I, and HpaII whereas both CpH/WpG methylation-sensitive enzymes (MspI and PspGI) identified around 90% of the annotated transposable elements, therefore showing almost no depletion. Furthermore, the covered area of the identified transposable elements for the three CpG methylation-sensitive enzymes is very low in comparison to the covered area of the gene models. The depletion of transposable elements is even stronger when only hypomethylated regions comprising at least five reads are considered. This restriction causes a depletion of identified transposable elements to 14% and at the same time on average 68% of the gene models were identified. For HpaII and AciI even 74% of the annotated gene models were identified.

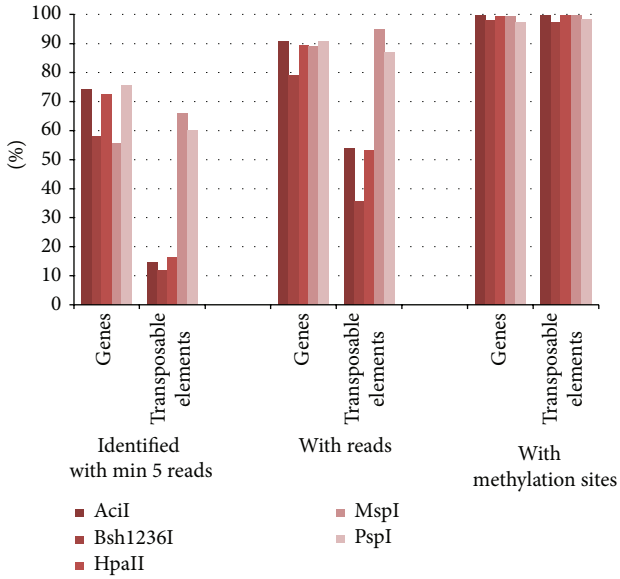


Figure 1: Genes and transposable elements identified in the rice genome with the methyl filtration technique. The regions comprised of at least five reads (left), and all regions (middle) show a clear depletion of transposable elements for AciI, Bsh1236I, and HpaII. On the right a representation of genes and transposable elements is given showing potential methylation sites within their gene space. All values are shown in percent based on the annotated 39.954 genes and 15.847 transposable elements.

Based on these results HpaII and AciI were selected as enzymes showing the best gene space coverage and the highest transposable element (TE) depletion.

3.2. Isolated Hypomethylated Fragments Are Preferentially Located in the Gene Space. The sequenced MF fragments were identified in the rice genome by mapping the high quality sequence reads to the genome sequence. The analysis was performed with a combination of both enzyme datasets, further referred to as combined dataset, and each dataset separately to investigate the complementarity of the two enzymes, AciI and HpaII. The combined dataset resulted in 129.810 regions representing 19% of the complete rice genomic sequence with an average coverage per bp of 61x integrating 80% of the reads. The single enzyme datasets identified 98.355 regions for AciI (coverage: 20x, reads: 69%) and 84.874 for HpaII (coverage: 40x, reads: 86%) representing 10% and 13% of the genome, respectively (see Figure 2 and Figure 3).

Figure 3 shows the overlap between the three datasets. All regions identified with the datasets of the single enzymes were also identified with the combined dataset. The comparison between both enzyme specific datasets shows the complementarity of both enzymes in respect of the identified genomic area. Whereas an overlap between both enzymes is clearly present, the majority of the identified genomic area is identified by only one of the enzyme datasets. This difference is mainly caused by the genomic location of the specific restriction sites of the used enzymes. Figure 4 depicts

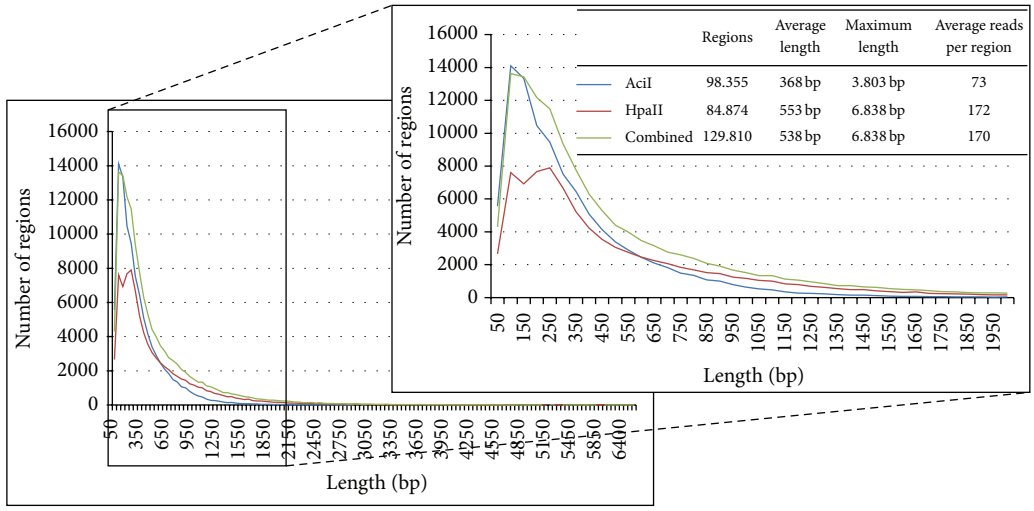


Figure 2: Length distribution of genomic regions identified for AciI, HpaII, and the combined dataset in rice. The length distribution of hypomethylated regions identified with the three datasets up to the maximal length is shown as well as a closer view to the region between 0 and 2.000 bp, where an increase in length is visible for the combined dataset. Additionally, the amount of regions, the average and maximum length, and the average reads per region are given.

	Genome area		Gene models		Genes		Transposable elements	
Total	373.245.519 bp		55.801		39.954		15.847	
Combined	69.808.680 bp	19%	37.247	67%	33.740	84%	3.507	22%
AciI	36.155.399 bp	10%	31.981	57%	29.671	74%	2.310	15%
HpaII	46.965.884 bp	13%	31.630	57%	29.029	73%	2.601	16%

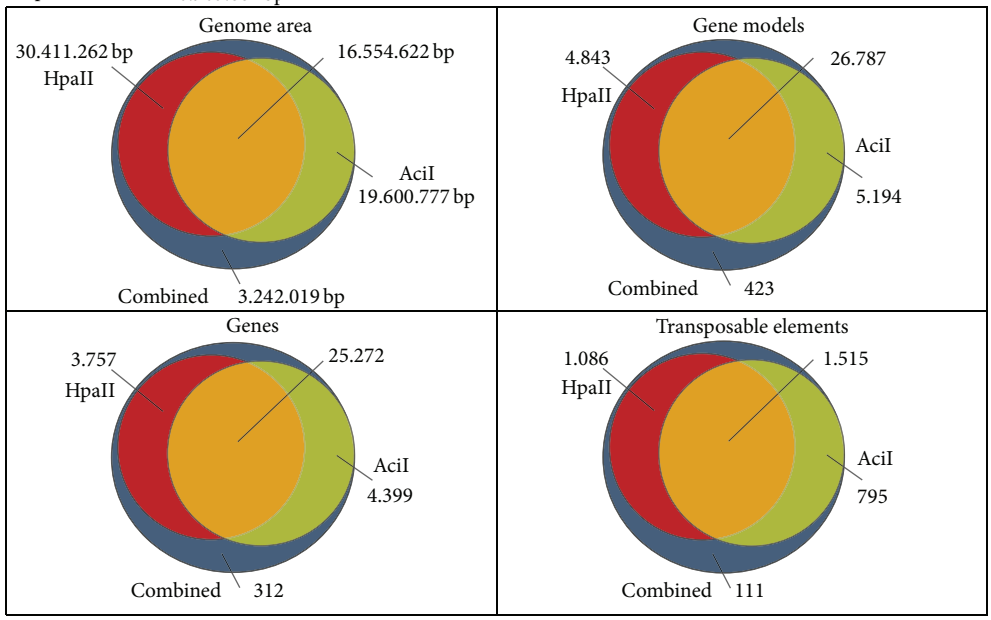


Figure 3: Overlap between the datasets in rice. The hypomethylated regions identified with the three datasets are compared focussing on genomic area, annotated gene models (including both genes and TEs and their surrounding +/− 1.000 bp regions), and genes and TEs separately. The total in each category is given in the table above while the overlap is visualized in the separate Venn diagrams.

an example of this scenario where distinct and overlapping regions for both enzymes were identified due to absence of the recognition sites for the other enzyme. The locations of the specific sites also affect the length of the regions, which is also reflected by their increased length in the combined dataset compared to the single datasets, especially in the range between 200 and 1.000 bp (Figure 2), showing an advantage of using the combination of both enzymes. In addition, the combined dataset identified extra 3 Mb of the genome which was not detected with only one dataset. These

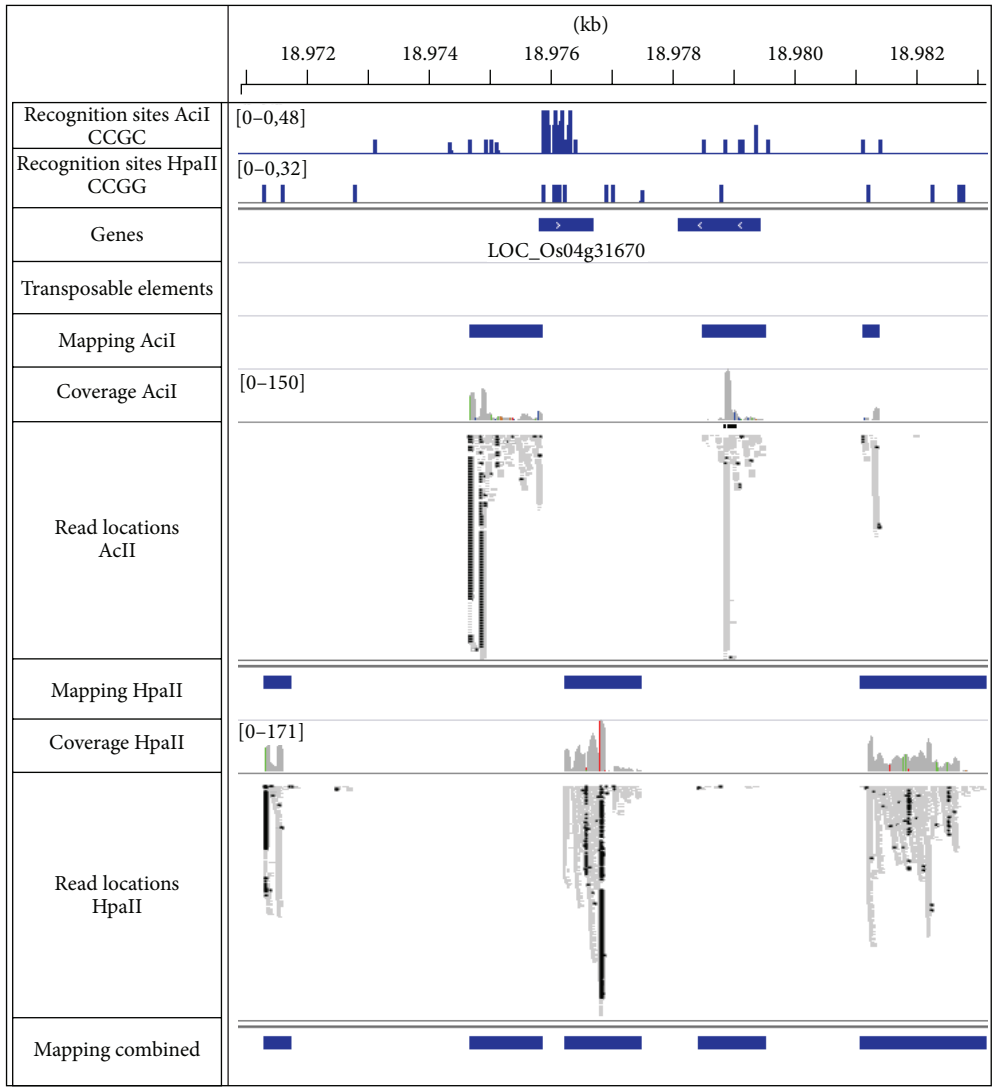


Figure 4: Complementary identification of genomic regions in rice due to restriction site locations. A detailed representation of the mapping results is shown for both enzymes, AciI and HpaII. The identified regions around the displayed gene differ due to the lack of recognition sites for the other enzyme. On the right, an example for overlapping but expanded regions is given.

additional regions contain reads from both single enzyme datasets but did not exceed the minimal coverage applied as quality insurance within each individually. Therefore they represent overlapping regions between both single enzyme datasets and could most likely also be identified separately by increasing the initial sequencing coverage per enzyme dataset.

Analysing the location of the regions within the genome, it was found that 67% of the regions in the combined dataset were identified either within a gene model (both genes and TEs) or in their flanking 1.000 bp area. Separating genes and TEs clearly shows the depletion of TEs and the enrichment for the gene space (gene body and flanking 1.000 bp; see Figure 3). Of the 39.954 annotated genes 84% were identified with the combined dataset, whereas of the annotated 15.847 TEs only 22% showed a hypomethylated region. This observation is also clear in both single datasets (74% genes and 15% TE for AciI, and 73% genes and 16% TE for HpaII; see Figure 3), which is also reflected in the distribution of the isolated fragments across the genome corresponding clearly with the locations of the genes and opposing the distribution of TEs (Figure 5).

Detailed analysis of the locations emphasized the preference for the gene space and especially the gene body. In the combined dataset 75% of the regions were located within the gene space. Of those 67% are either located within or overlapping exons. This preference is even more prominent regarding the genomic area. 78% of the genomic area identified with the combined dataset is annotated as gene space and thereof 76% is associated with exons (see Table 1). Besides the enrichment for exons, another 16% of the isolated regions that are located in the gene space represent parts of the upstream 1.000 bp regions, of which 57% show known transcription factor binding sites.

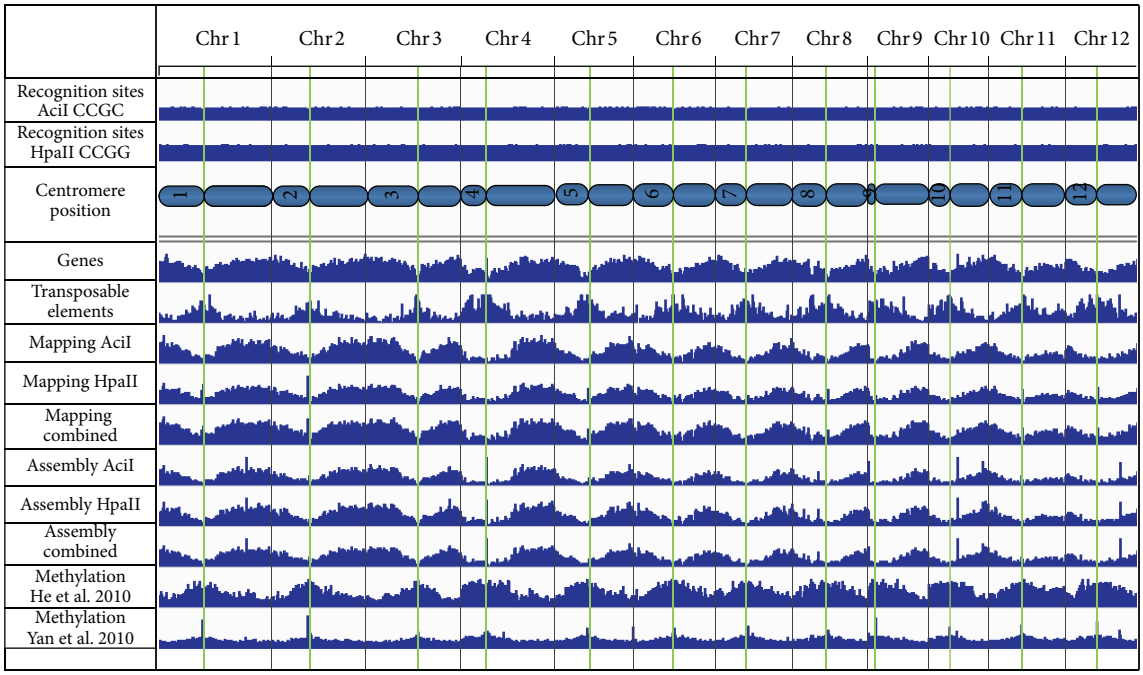


Figure 5: Genomic overview of hypomethylated regions in rice. The results of the *de novo* assembly and mapping approach are displayed for all three datasets (AciI, HpaII, and combined). In the upper panel the positions of the recognition sites are shown. The locations of annotated genes and TEs are depicted as well as the regions identified with the data of previous methylation studies (lower two tracks). The positions of the centromeric regions are also indicated and represented by green lines.

Table 1: Allocation of the identified hypomethylated regions to the gene space in rice: the position of the identified hypomethylated regions is given in respect of the gene space together with the genomic area.

	AciI		HpaII		Combined	
Hypomethylated regions						
Total	98.355		84.874		129.810	
Gene space (±1.000 bp)	76.295	78%	62.655	74%	96.717	75%
Upstream 1.000 bp	13.342	17%	10.876	17%	15.629	16%
Downstream 1.000 bp	7.739	10%	6.398	10%	9.556	10%
Gene body	55.214	72%	45.381	72%	71.532	74%
Exon	48.933	64%	41.034	65%	64.622	67%
Intron	6.281	8%	4.347	7%	6.910	7%
Genomic area [bp]						
Total	36.155.399		46.965.884		69.808.680	
Gene space (±1.000 bp)	28.580.311	79%	36.521.506	78%	54.732.127	78%
Upstream 1.000 bp	4.439.728	16%	5.111.923	14%	6.931.498	13%
Downstream 1.000 bp	2.650.795	9%	2.949.893	8%	4.235.613	8%
Gene body	21.489.788	75%	28.459.690	78%	43.565.016	80%
Exon	19.739.225	69%	27.027.474	74%	41.407.771	76%
Intron	1.750.563	6%	1.432.216	4%	2.157.245	4%

Further comparative analysis with previously published results showed that 20% of the identified upstream area corresponds to regions of open chromatin in rice seedlings [53] indicating a potential active state of promoter elements in gene regulation. The isolation of hypomethylated regions via MF does therefore not only provide a representation of coding sequences within the gene space but does also provide potentially active regulatory upstream regions.

3.3. Sequencing Coverage Simulation to Identify the Hypomethylome of the Whole Genome. The overall sequence coverage was estimated based on the size of the complete genome sequence as 7x for the combined dataset, 3x for AciI, and 5x for HpaII, respectively, providing a reliable base for the study of the hypomethylated regions of the complete genome. For estimating the minimal coverage necessary to identify the hypomethylome of the whole genome, a simulation was

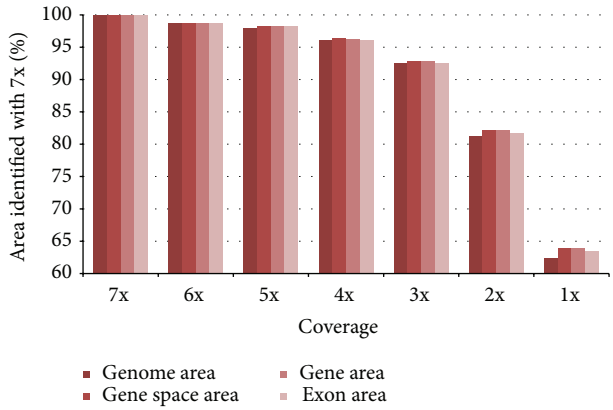


Figure 6: A simulation performed in rice to estimate the minimal coverage necessary to identify the hypomethylome of the whole genome was performed by randomly selecting reads from the combined dataset with 7x coverage to represent different coverage thresholds. The reads were allocated to the genome sequence and compared to the result of the complete dataset (100%) regarding genome area, gene space area, gene area, and exon area.

performed by randomly selecting reads from the complete dataset to represent different coverage thresholds (2–7x) and remapping these to the genome sequence. The recovery of the allocated regions of the simulation datasets with the complete dataset was calculated together with the identified genomic area and the location within the gene space, gene body, and exon area, respectively. The data showed only a slight decrease in identified genomic area for 6x and 5x coverage. Also using a coverage of 4x results in 96% of the identified genomic area. But a clear decrease to 81% compared to the complete dataset can be observed with a coverage of 2x, while with a coverage of 3x still 92% were identified (Figure 6). The decrease of gene space, gene body, and exon area showed a very similar distribution.

Therefore we recommend a minimal coverage of 3-4x for similar studies. Hence, both single enzyme datasets also show enough coverage to represent the hypomethylome of the whole rice genome.

3.4. De Novo Assembled Contigs of Hypomethylated Regions Are Highly Similar to the Results of the Reference Based Identification. The *de novo* assembly of the datasets resulted in 187.168 contigs for the combined dataset integrating 82% of the reads, 129.402 contigs for the enzyme AciI (reads: 70%), and 111.390 contigs for HpaII (reads: 87%). The assembled contigs were located in the rice genome, where 95% could be identified. The contigs which could not be located within the genomic sequence consist mainly of the additional reads (2%) that could be assembled but could not be located with the mapping approach. A similarity search to the NT database of NCBI indicated no genic origin for these contigs suggesting a nongenic origin of the respective fragment with lower evolutionary selection pressure. Hence these regions might show more differences to the reference genome and could therefore not be identified based on similarity thresholds.

The genomic area represented by the *de novo* assembled contigs overlapped to 95% with the genomic area identified by the mapping approach. One fifth of the nonoverlapping genomic area was gained by the slightly longer regions produced by the assembly approach. However, the majority (about 80%) is mainly located in distinct regions close to the centromere or genomic chloroplast and mitochondrial regions as depicted in Figure 5. The *de novo* assembled contigs which could be located in these genomic chloroplast and mitochondrial regions showed differences to the plastid genomes. Therefore these contigs most likely do represent genomic hypomethylated regions, especially as all reads which showed high similarity to the plastid sequences of the chloroplast and the mitochondrion were filtered in the initial preprocessing. However they show also enough sequence differences to the published genomic sequence to prevent their identification with the mapping approach, indicating sequence variations between individuals which could hint to regions under lower selective pressure.

However these differences originate mainly from the HpaII dataset. The enzymes HpaII and AciI cleave only at a potential cut site if nonmethylated cytosines are present, therefore enriching hypomethylated regions. The observed difference in the two datasets would suggest the differential methylation of the recognition sites of the two enzymes, since these regions show evenly distributed recognition sites for both enzymes.

Comparing the hypomethylated regions identified in our system with the data of Yan et al. [14] representing methylated regions obtained by immunoprecipitation, we found 3% overlap. This overlap is mainly located in the regions identified only by HpaII. This contradicting information also suggests a different kind of methylation in these regions which does not prevent HpaII to cleave at its recognition site, while still preventing cleavage by AciI. The observed small overlap between the hypomethylated regions and methylated areas is also confirmed in the comparison to other datasets (Figure 5). A comparison of our hypomethylated regions with the methylation study of He et al. [51] indicated methylation in 8% of the genomic area identified as hypomethylated in our study. Those regions might represent areas which are differentially methylated in different individuals, developmental stages, or tissues. Similar results have recently also been shown in maize [54].

Our results of 19% identified hypomethylated regions in the rice genome are furthermore in good agreement with data of a previous study stating 76–91% genome coverage of methylated regions [52].

3.5. Applicability of the Method to Other Genomes. The nearly identical results of both the reference based mapping and the *de novo* assembly demonstrate that a reliable representation of the hypomethylated regions in a genome can be identified not only if a reference genome sequence is available but also by applying a *de novo* assembly approach. The applicability of the method is however dependent on the frequency of the recognition sites of the applied enzymes, which is in rice 6.2 sites/kb for AciI and 2.8 sites/kb for HpaII. An *in silico* analysis of several fully sequenced genomes showed

a similar frequency in monocotyledon plants (AciI: maize 5.7 sites/kb, sorghum 4.1 sites/kb, *Brachypodium* 6.5 sites/kb; HpaII: maize 3.3 sites/kb, sorghum 2.2 sites/kb, *Brachypodium* 3.5 sites/kb). In other angiosperms and gymnosperms the frequencies are less but similar to frequencies observed in the human genome, where these enzymes are also used to study genome wide methylation patterns [21] (AciI: human 1.1 sites/kb, *Arabidopsis* 1.7 sites/kb, poplar 1.0 sites/kb, grapevine 0.9 sites/kb, Norway spruce 0.9 sites/kb; HpaII: human 0.8 sites/kb, *Arabidopsis* 1.1 sites/kb, poplar 0.7 sites/kb, grapevine 0.7 sites/kb, Norway spruce 0.5 sites/kb). This renders the presented technique highly applicable for nonmodel organisms where no genome sequence is available.

3.6. Applicability to the Large Genome of Norway Spruce. Despite the advances in sequencing technologies and the still increasing amount of sequenced genomes in the last decade one challenging issue remains especially for large plant genomes: their high amount of repetitive regions in the genome sequences. Genome sizes in plants range from 63 Mb up to about 150 Gb [55–57]. While gene size and number are rather constant with 30.000–50.000 genes, the differences in genome size are mainly due to the abundance of repetitive DNA which represents the majority of the genomic sequence [33, 36, 58, 59]. For example, the 20 Gb nuclear genome of Norway spruce contains 70% high copy number repeats and only about 2.4% of the nuclear genomic sequence, characterized as genes or gene-like fragments (possible pseudogenes) [33]. Repetitive regions have proven to be the main challenge in genome assembly approaches especially when using whole genome shotgun approaches. It has been suggested that silencing instead of chromosome rearrangement is the predominant mechanism in Norway spruce to deal with the high amount of repetitive regions [33]. One of the primary mechanisms that cause gene silencing is the methylation of DNA [4, 16], which renders Norway spruce a good example to emphasize the applicability of our technique to large genomes. Our technique was applied to isolate hypomethylated fragments from different samples of Norway spruce using only HpaII. Additionally non-filtrated genomic DNA was sequenced as comparison. Each dataset was treated and mapped to the published genomic sequence as described for the rice datasets. The resulting sequence reads represent an average coverage of the Norway spruce genome of 0.2x in the four samples and about 1x for the random genomic sequences.

Due to the lower coverage in these datasets we do not expect to gain the complete hypomethylome, although we observe the same advantages of depletion of transposable elements and the enrichment for the gene space. Particularly the comparison to the non-filtrated genomic dataset highlights both effects of our technique (Figure 7). This clear depletion and the similarity of the results to the rice results show that our approach is applicable also for the large Norway spruce genome and is not affected by the different composition of the genome or other repeat-classes as it depends on the DNA methylation pattern in the genome. However, as has been seen in the rice data, different methylation pattern can influence the detection. These results show the applicability of the technique to large genomes.

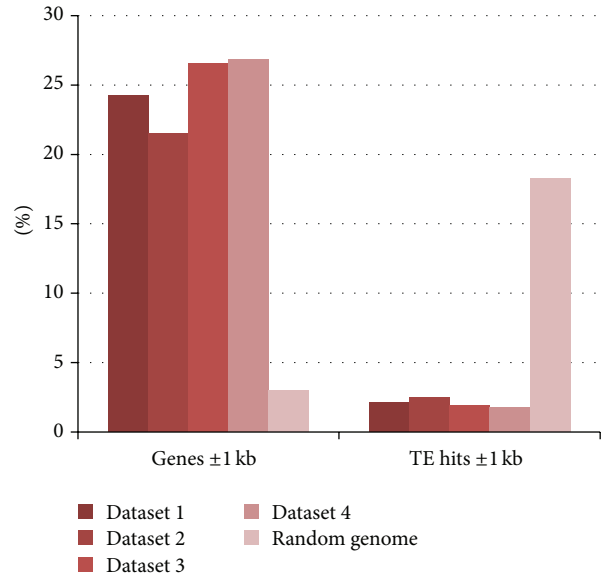


Figure 7: Reads located within the gene space (gene and surrounding +/− 1.000 bp regions) and the annotated TE (including surrounding +/− 1.000 bp regions) in Norway spruce. A clear enrichment of reads derived from gene regions and a clear depletion of reads derived from TE regions are shown.

3.7. Applicability to Polyploid Genomes. Additionally, our technique has been applied to the triploid saffron crocus genome using only HpaII on one sample. The resulting reads have been preprocessed and *de novo* assembled as described for the rice datasets. The genome of saffron crocus has so far not been sequenced mainly due to its complexity. Its genome was estimated to have a size of 10.3 Gb [60]. In addition to its size it is a triploid genome most likely derived through crossing between two closely related species [60–62], introducing the complexity of polyploidy and different allelic variants to the analysis which could affect the quality and reliability of the assembly and the gene space detection. However, we did not observe a decreased quality of the *de novo* assembly as the analysis resulted in 13.986 contigs integrating 81% of the sequence reads. Sequence comparison to other monocotyledon plants showed a similar depletion of transposable elements (1%) and a similar enrichment for genes as 26% of the assembled contigs show similarity to protein coding genes. The identification of regions located in the flanking regions is not directly possible as no reference genome is available; however, 47% of the contigs show known regulatory elements, indicating the isolation of active regulatory regions. These results indicate that with our approach the detection of the gene space including regulatory regions is not affected when applied to polyploid genomes.

Although the overall genome coverage is rather small (0.2x), the obtained results are comparable to the data of Norway spruce and rice since a similar enrichment for

the gene space and at the same time a clear depletion of transposable elements was detected. The data obtained on saffron crocus gave further support for the universal applicability of the presented method on a wide range of plant genomes, including also complex polyploid genomes.

4. Conclusions

Because of the large size and high complexity of many plant genomes, particularly those of important crops, gene-enriched sequencing strategies have been designed as an alternative to whole genome sequencing in an attempt to capture the gene space (genes plus regulatory elements) of such genomes. One of these enrichment techniques, called methyl filtration (MF), takes advantage of the difference in methylation state of cytosine residues being present between the gene space and repetitive elements.

With the present study performed on the model organism rice we demonstrate that with an improved MRE-seq method followed by *de novo* assembly, the hypomethylome and thus the gene space of a plant can be easily accessed. Using two methylation-sensitive restriction enzymes (HpaII and AciI), 84% of the annotated coding regions of the rice genome could be isolated, meanwhile, reducing the amount of isolated transposable elements to one fifth. The latter is of utmost importance to enrich the gene space as most plant genomes consist of 50–80% of repetitive elements including TEs. The presented method filters the genes including exons and introns as well as their up- and downstream flanking regions where regulatory elements are located. This represents a clear advantage over traditional transcriptome analysis approaches, which provide sequence data only of exons of active genes.

De novo assembly shows almost identical results as reference based mapping of the sequence reads, demonstrating the applicability of the MRE-seq approach to nonmodel plant species where no fully sequenced genome is available. The coverage needed for generating an informative snapshot of a given genome is estimated as 3-4x. Small scale studies in the large genome of Norway spruce and the polyploid saffron crocus demonstrate the depletion of transposable elements and enrichment for the gene space in nonmodel species and complex plant genomes.

The overlap of our results with methylation data from previous studies confirms the high potential of MRE-seq for being applied in a wide range of scenarios for the direct analysis of methylation differences, for example, between ecotypes and individuals, and within and across species. This new and easy technique allows the fast and inexpensive generation of data necessary for studying the role of epigenetics as a source of adaptive variation in natural populations as well as crop plants. It is especially helpful with regard to studying and understanding the genomes of nonmodel organisms.

Conflict of Interests

The authors declare that there is no conflict of interests related to the data presented in this publication.

Authors' Contribution

Kornel Burg, Maria Berenyi, and Silvia Fluch designed the experiment. Maria Berenyi and Karin Hansel-Hohl performed sample collection and DNA extraction as well as wet lab methyl filtration experiments. Elisabeth Wischnitzki performed bioinformatics analyses. Silvia Fluch and Kornel Burg advised in experimental design and assisted in conceptual data analysis. Elisabeth Wischnitzki and Eva Maria Sehr drafted the paper. Silvia Fluch coordinated the project and edited the paper. All the authors read and approved the final version of the paper.

Acknowledgment

The authors wish to thank the AIT Austrian Institute of Technology GmbH for financial support.

References

[1] M. Y. Kim and D. Zilberman, "DNA methylation as a system of plant genomic immunity," *Trends in Plant Science*, vol. 19, no. 5, pp. 320–326, 2014.

[2] R. Gutzat and O. Mittelsten Scheid, "Epigenetic responses to stress: triple defense?" *Current Opinion in Plant Biology*, vol. 15, no. 5, pp. 568–573, 2012.

[3] T. J. Hardcastle, "High-throughput sequencing of cytosine methylation in plant DNA," *Plant Methods*, vol. 9, no. 1, article 16, 2013.

[4] P. D. Rabinowicz, R. Citek, M. A. Budiman et al., "Differential methylation of genes and repeats in land plants," *Genome Research*, vol. 15, no. 10, pp. 1431–1440, 2005.

[5] J. Wang, N. C. Marowsky, and C. Fan, "Divergence of gene body DNA methylation and evolution of plant duplicate genes," *PLoS ONE*, vol. 9, no. 10, Article ID e110357, 2014.

[6] F. M. Piccolo and A. G. Fisher, "Getting rid of DNA methylation," *Trends in Cell Biology*, vol. 24, no. 2, pp. 136–143, 2014.

[7] N. Widman, S. Feng, S. E. Jacobsen, and M. Pellegrini, "Epigenetic differences between shoots and roots in *Arabidopsis* reveals tissue-specific regulation," *Epigenetics*, vol. 9, no. 2, pp. 236–242, 2014.

[8] Y.-Y. Zhang, M. Fischer, V. Colot, and O. Bossdorf, "Epigenetic variation creates potential for evolution of plant phenotypic plasticity," *New Phytologist*, vol. 197, no. 1, pp. 314–322, 2013.

[9] K. D. Kim, M. El Baidouri, and S. A. Jackson, "Accessing epigenetic variation in the plant methylome," *Briefings in Functional Genomics*, vol. 13, no. 4, pp. 318–327, 2014.

[10] N. Plongthongkum, D. H. Diep, and K. Zhang, "Advances in the profiling of DNA modifications: cytosine methylation and beyond," *Nature Reviews Genetics*, vol. 15, no. 10, pp. 647–661, 2014.

[11] F. Mohn, M. Weber, D. Schübeler, and T.-C. Roloff, "Methylated DNA immunoprecipitation (MeDIP)," in *DNA Methylation*, J. Tost, Ed., pp. 55–64, Humana Press, 2009.

[12] M. Gao, Q. Huang, Y. Chu, C. Ding, B. Zhang, and X. Su, "Analysis of the leaf methylomes of parents and their hybrids provides new insight into hybrid vigor in *Populus deltoides*," *BMC Genetics*, vol. 15, supplement 1, article S8, 2014.

[13] S. C. Stelpflug, S. R. Eichten, P. J. Hermanson, N. M. Springer, and S. M. Kaeppler, "Consistent and heritable alterations of

DNA methylation are induced by tissue culture in maize," *Genetics*, vol. 198, no. 1, pp. 209–218, 2014.

[14] H. Yan, S. Kikuchi, P. Neumann et al., "Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice," *Plant Journal*, vol. 63, no. 3, pp. 353–365, 2010.

[15] L. E. Palmer, P. D. Rabinowicz, A. L. O'Shaughnessy et al., "Maize genome sequencing by methylation filtration," *Science*, vol. 302, no. 5653, pp. 2115–2117, 2003.

[16] P. D. Rabinowicz, L. E. Palmer, B. P. May et al., "Genes and transposons are differentially methylated in plants, but not in mammals," *Genome Research*, vol. 13, no. 12, pp. 2658–2664, 2003.

[17] E. A. Raleigh, N. E. Murray, H. Revel et al., "McrA and McrB restriction phenotypes of some *E. coli* strains and implications for gene cloning," *Nucleic Acids Research*, vol. 16, no. 4, pp. 1563–1575, 1988.

[18] C. A. Whitelaw, W. B. Barbazuk, G. Pertea et al., "Enrichment of gene-coding sequences in maize by genome filtration," *Science*, vol. 302, no. 5653, pp. 2118–2120, 2003.

[19] B. Zhang, X. Xing, J. Li et al., "Comparative DNA methylome analysis of endometrial carcinoma reveals complex and distinct deregulation of cancer promoters and enhancers," *BMC Genomics*, vol. 15, no. 1, article 868, 2014.

[20] A. K. Maunakea, R. P. Nagarajan, M. Bilenky et al., "Conserved role of intragenic DNA methylation in regulating alternative promoters," *Nature*, vol. 466, no. 7303, pp. 253–257, 2010.

[21] D. Li, B. Zhang, X. Xing, and T. Wang, "Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation," *Methods*, vol. 72, pp. 29–40, 2015.

[22] M. Berenyi, R. P. Mauleon, D. Kopecky et al., "Isolation of plant gene space-related sequence elements by high C+G patch (HCGP) filtration: model study on rice," *Plant Molecular Biology Reporter*, vol. 27, no. 1, pp. 79–85, 2009.

[23] H. Sakai, H. Ikawa, T. Tanaka et al., "Distinct evolutionary patterns of *Oryza glaberrima* deciphered by genome sequencing and comparative analysis," *Plant Journal*, vol. 66, no. 5, pp. 796–805, 2011.

[24] C. Grativol, M. Regulski, M. Bertalan et al., "Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus *Saccharum*," *The Plant Journal*, vol. 79, no. 1, pp. 162–172, 2014.

[25] E.-T. L. Low, R. Rosli, N. Jayanthi et al., "Analyses of hypomethylated oil palm gene space," *PLoS ONE*, vol. 9, no. 1, Article ID e86728, 2014.

[26] Gayacharan and A. J. Joel, "Epigenetic responses to drought stress in rice (*Oryza sativa* L.)," *Physiology and Molecular Biology of Plants*, vol. 19, no. 3, pp. 379–387, 2013.

[27] X. Chen, J. Hu, H. Zhang, and Y. Ding, "DNA methylation changes in photoperiod-thermo-sensitive male sterile rice PA64S under two different conditions," *Gene*, vol. 537, no. 1, pp. 143–148, 2014.

[28] C. Hébrard, M.-V. Trap-Gentil, C. Lafon-Placette et al., "Identification of differentially methylated regions during vernalization revealed a role for RNA methyltransferases in bolting," *Journal of Experimental Botany*, vol. 64, no. 2, pp. 651–663, 2013.

[29] X. Ou, Y. Zhang, C. Xu et al., "Transgenerational inheritance of modified DNA methylation patterns and enhanced tolerance induced by heavy metal stress in rice(*Oryza sativa* L.)," *PLoS ONE*, vol. 7, no. 9, Article ID e41143, 2012.

[30] B. Piegu, R. Guyot, N. Picault et al., "Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice," *Genome Research*, vol. 16, no. 10, pp. 1262–1269, 2006.

[31] T. Nussbaumer, M. M. Martis, S. K. Roessner et al., "MIPS PlantsDB: a database framework for comparative plant genome research," *Nucleic Acids Research*, vol. 41, no. 1, pp. D1144–D1151, 2013.

[32] Y. Kawahara, M. de la Bastide, J. P. Hamilton et al., "Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, article 1, 2013.

[33] B. Nystedt, N. R. Street, A. Wetterbom et al., "The Norway spruce genome sequence and conifer genome evolution," *Nature*, vol. 497, no. 7451, pp. 579–584, 2013.

[34] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[35] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[36] Arabidopsis Genome Initiative, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*," *Nature*, vol. 408, no. 6814, pp. 796–815, 2000.

[37] G. A. Tuskan, S. DiFazio, S. Jansson et al., "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)," *Science*, vol. 313, no. 5793, pp. 1596–1604, 2006.

[38] O. Jaillon, J.-M. Aury, B. Noel et al., "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla," *Nature*, vol. 449, no. 7161, pp. 463–467, 2007.

[39] P. S. Schnable, D. Ware, R. S. Fulton et al., "The B73 maize genome: complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, 2009.

[40] A. H. Paterson, J. E. Bowers, R. Bruggmann et al., "The *Sorghum bicolor* genome and the diversification of grasses," *Nature*, vol. 457, no. 7229, pp. 551–556, 2009.

[41] International Brachypodium Initiative, "Genome sequencing and analysis of the model grass *Brachypodium distachyon*," *Nature*, vol. 463, no. 7282, pp. 763–768, 2010.

[42] R. Hehl and L. Bülow, "AthaMap web tools for the analysis of transcriptional and posttranscriptional regulation of gene expression in *Arabidopsis thaliana*," in *Plant Circadian Networks*, vol. 1158 of *Methods in Molecular Biology*, pp. 139–156, Springer, New York, NY, USA, 2014.

[43] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga, "Plant cis-acting regulatory DNA elements (PLACE) database: 1999," *Nucleic Acids Research*, vol. 27, no. 1, pp. 297–300, 1999.

[44] A. Mathelier, X. Zhao, A. W. Zhang et al., "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. 42, no. 1, pp. D142–D147, 2014.

[45] V. Matys, E. Fricke, R. Geffers et al., "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003.

[46] A. Yilmaz, M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold, "AGRIS: the arabidopsis gene regulatory information server, an update," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1118–D1122, 2011.

[47] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[48] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[49] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler et al., "Integrative genomics viewer," *Nature Biotechnology*, vol. 29, no. 1, pp. 24–26, 2011.

[50] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 178–192, 2013.

[51] G. He, X. Zhu, A. A. Elling et al., "Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids," *Plant Cell*, vol. 22, no. 1, pp. 17–33, 2010.

[52] X. Li, J. Zhu, F. Hu et al., "Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression," *BMC Genomics*, vol. 13, no. 1, article 300, 2012.

[53] W. Zhang, Y. Wu, J. C. Schnable et al., "High-resolution mapping of open chromatin in the rice genome," *Genome Research*, vol. 22, no. 1, pp. 151–162, 2012.

[54] M. Lauria, S. Piccinini, R. Pirona, G. Lund, A. Viotti, and M. Motto, "Epigenetic variation, inheritance, and parent-of-origin effects of cytosine methylation in maize (*Zea mays*)," *Genetics*, vol. 196, no. 3, pp. 653–666, 2014.

[55] M. Bennett and I. J. Leitch, "Plant DNA C-values Database (Release 6.0, Dec. 2012)," 2012, http://data.kew.org/cvalues/.

[56] J. Greilhuber, T. Borsch, K. Müller, A. Worberg, S. Porembski, and W. Barthlott, "Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size," *Plant Biology*, vol. 8, no. 6, pp. 770–777, 2006.

[57] J. Pellicer, M. F. Fay, and I. J. Leitch, "The largest eukaryotic genome of them all?" *Botanical Journal of the Linnean Society*, vol. 164, no. 1, pp. 10–15, 2010.

[58] R. B. Flavell, M. D. Bennett, J. B. Smith, and D. B. Smith, "Genome size and the proportion of repeated nucleotide sequence DNA in plants," *Biochemical Genetics*, vol. 12, no. 4, pp. 257–269, 1974.

[59] S. Ouyang, W. Zhu, J. Hamilton et al., "The TIGR Rice Genome Annotation Resource: improvements and new features," *Nucleic Acids Research*, vol. 35, no. 1, pp. D883–D887, 2007.

[60] F. Brandizzi and M. G. Caiola, "Flow cytometric analysis of nuclear DNA in Crocus sativus and allies (Iridaceae)," *Plant Systematics and Evolution*, vol. 211, no. 3-4, pp. 149–154, 1998.

[61] M. G. Caiola, "Saffron reproductive biology," *Acta Horticulturae*, vol. 650, pp. 25–37, 2004.

[62] G. Frizzi, M. Miranda, C. Pantani, and F. Tammaro, "Allozyme differentiation in four species of the *Crocus cartwrightianus* group and in cultivated saffron (*Crocus sativus*)," *Biochemical Systematics and Ecology*, vol. 35, no. 12, pp. 859–868, 2007.