INFORMATICS

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

RRBS-Analyser: A Comprehensive Web Server for Reduced Representation Bisulfite Sequencing Data Analysis

Tao Wang,[1†] Qi Liu,[1†] Xianfeng Li,[1] Xiaobing Wang,[2] Jinchen Li,[1] Xiaochun Zhu,[2] Zhong Sheng Sun,[1,3*] and Jinyu Wu[1,4‡]

[1] *Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou 325035, China;* [2] *Division of Rheumatology, First Affiliated Hospital, Wenzhou Medical University, Wenzhou 325000, China;* [3] *Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China;* [4] *Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei 230026, China*

Communicated by Ulf Landegren

Received 28 June 2013; accepted revised manuscript 9 September 2013.
Published online 18 September 2013 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22444

ABSTRACT: In reduced representation bisulfite sequencing (RRBS), genomic DNA is digested with the restriction enzyme and then subjected to next-generation sequencing, which enables detection and quantification of DNA methylation at whole-genome scale with low cost. However, the data processing, interpretation, and analysis of the huge amounts of data generated pose a bioinformatics challenge. We developed RRBS-Analyser, a comprehensive genome-scale DNA methylation analysis server based on RRBS data. RRBS-Analyser can assess sequencing quality, generate detailed statistical information, align the bisulfite-treated short reads to reference genome, identify and annotate the methylcytosines (5mCs) and associate them with different genomic features in CG, CHG, and CHH content. RRBS-Analyser supports detection, annotation, and visualization of differentially methylated regions (DMRs) for multiple samples from nine reference organisms. Moreover, RRBS-Analyser provides researchers with detailed annotation of DMR-containing genes, which will greatly aid subsequent studies. The input of RRBS-Analyser can be raw FASTQ reads, generic SAM format, or self-defined format containing individual 5mC sites. RRBS-Analyser can be widely used by researchers wanting to unravel the complexities of DNA methylome in the epigenetic community. RRBS-Analyser is freely available at http://122.228.158.106/RRBSAnalyser/.**
Hum Mutat 34:1606–1610, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: methylation; next-generation sequencing; RRBS; differentially methylated region; epigenetics

Introduction

DNA methylation has important functions in the regulation of gene expression during various biological processes, such as X chromosome inactivation, genomic imprinting, embryogenesis, and maintenance of genomic integrity [Bird, 2002; Harris et al., 2010]. Aberrations in DNA methylation have been implicated in many diseases and traits, such as autoimmune disorders, aging, and cancer. Typically, DNA methylation occurs at the 5′-carbon position of cytosine within a CpG dinucleotide in plants and mammals, although it also occurs at CHH and CHG cytosines. For decades, the gold standard for DNA methylation analysis has been bisulfite sequencing based on traditional Sanger sequencing [Warnecke et al., 2002]. Bisulfite treatment of DNA, followed by PCR amplification, leads to a chemical conversion of unmethylated Cs to Ts, while leaving methylated Cs unchanged [Frommer et al., 1992]. However, this procedure is very laborious and time-consuming, and is, therefore, inappropriate for high-throughput studies.

Currently, next-generation sequencing (NGS) has been widely applied to characterize DNA methylation, because of its capacity to generate massive amounts of data in a short time, which provides an unprecedented opportunity to discover DNA methylation sites on a genome-wide scale (DNA methylome) [Ku et al., 2011]. RRBS (reduced representation bisulfite sequencing), which combines NGS, bisulfite conversion, and restriction enzyme digestion, is an efficient method for investigating DNA methylation at single-nucleotide resolution with higher efficiency and lower cost in comparison with whole-genome bisulfite sequencing [Meissner et al., 2005]. It enriches genome areas with a high CpG content, which greatly reduces the sample DNA required. The improved bisulfite treatment protocol of RRBS also optimizes the conversion of unmethylated cytosines and minimizes the DNA loss due to bisulfite-induced degradation. Therefore, it is highly sensitive and provides quantitative DNA methylation measurements. However, the massive amount of data generated by NGS poses a great bioinformatics challenge in terms of data processing and analysis. Recently, a number of computational methods have been developed for mining the DNA methylation data generated by NGS, such as RRBSMAP [Xi et al., 2012], BS Seeker [Chen et al., 2010], Bismark [Krueger and Andrews, 2011], PASH [Coarfa et al., 2010], RMAP [Smith et al., 2009], BRAT-bw [Harris et al., 2012], SAAP-RRBS [Sun et al., 2012], methylKit [Akalin et al., 2012], Meth Tools 2.0 [Grunau et al., 2000], MethylAnalyzer [Xin et al., 2011], BSmooth [Hansen et al., 2012], EpiExplorer [Halachev et al., 2012], GBSA [Benoukraf et al., 2013], and QDMRs [Zhang et al., 2011]. Among them, RRBSMAP is a

†These authors contributed equally to this work.

‡Correspondence to: Jinyu Wu, Xueyuan West Road, Lu Cheng District, Wenzhou 325035, China. E-mail: iamwujy@gmail.com

*Correspondence to: Zhong Sheng Sun, Beichen West Road, Chao Yang District, Beijing 100101, China. E-mail: sunzs@mail.biols.ac.cn

Contract grant sponsors: National Natural Science Foundation of China (31171236/C060503); China–Canada Collaboration Project from Ministry of Science and Technology of China (2011DFA30670); National High Technology Research and Development Program of China (2012AA02A201, 2012AA02A202); Key Science and Technology Innovation Team of Zhejiang Province (2012R10048-05).

short-read alignment tool for handling RRBS data [Xi et al., 2012]. It uses wildcard alignment to enhance the computational efficiency of large-scale epigenome association studies performed with RRBS. SAAP-RRBS integrates read quality assessment, alignment, methylation data extraction, annotation, and visualization of RRBS data [Sun et al., 2012]. methylKit is an R package that performs clustering, sample quality visualization, differential methylation analysis, and annotation for both whole-genome data and RRBS data [Akalin et al., 2012].

In this study, an integrated platform, RRBS-Analyser, is developed to support comprehensive DNA methylation analyses in multiple organisms, which supports more functionality compared with the tools mentioned above. First, RRBS-Analyser provides a detailed assessment of the quality of the short reads, including CG dinucleotide distribution, distribution of reads with varied GC content, and length distribution of clean reads. Second, RRBS-Analyser detects and annotates methylcytosines (5mCs), showing global alignment information; detailed annotation of 5mC in CG, CHG, and CHH nucleotides; and bisulfite conversion rates. Furthermore, RRBS-Analyser supports multiple sample analyses, which is convenient for flexible differentially methylated region (DMR) detection and annotation. Additionally, RRBS-Analyser provides researchers with detailed annotation of DMR-containing genes, which will greatly aid subsequent experimental or bioinformatics studies.

RRBS-Analyser Analysis Workflow

Bisulfite-Treated Short Reads Quality Assessment

The procedure used by RRBS-Analyser to analyze the DNA methylome from RRBS data is detailed below. First, RRBS-Analyser filters low-quality reads using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Then, 3′/5′ adapters are trimmed using Cutadapt (http://code.google.com/p/cutadapt/) implemented in Trim Galore. The remaining data are defined as clean reads generated after removing adapters and low-quality base from raw reads. Last, FastQC (http://galaxy.csdb.cn:8000/tool_runner?tool_id=fastqc) is used to display quality information for clean reads.

5mCs Detection and Annotation

RRBS-Analyser employs RRBSMAP [Xi et al., 2012] for bisulfite-treated short reads alignment, and the binomial test with false discovery rate (FDR) constraint to identify the position of 5mCs from the alignment results file [Li et al., 2010; Lister et al., 2009]. During this procedure, the most frequent problems encountered that affect the detection of methylation are incomplete bisulfite conversion of cytosines and sequencing errors. To improve the accuracy of 5mCs identification, the minimum sequence depth threshold is determined at a cytosine position by the binomial probability distribution. Once the procedure of identified 5mC is completed, RRBS-Analyser creates a file to store the basic statistical information, including mapping information, methylation information, and bisulfite conversion rate. The bisulfite conversion rate is estimated through non-CG methylation status [Li et al., 2010]. In addition, RRBS-Analyser displays distribution of sequencing depth and restriction fragment lengths.

Subsequently, RRBS-Analyser displays detailed methylation information on different gene regions, such as 3′-UTR, 5′-UTR, CDS, introns, promoter (1,200 bp upstream and 300 bp downstream of the transcriptional start site), and repeat elements (such as LINE, SINE, satellite, simple repeat, and LTR), CpG islands, and intergenic regions. These annotations are downloaded from UCSC (http://genome.ucsc.edu/). Annotation of 5mCs is based on the association of the chromosome coordinates of 5mCs with the corresponding genomic annotation information.

DMR Detection and Annotation

The sliding window method is used to identify DMRs with a defined window size (default 200 bp) and a defined step size (default 10 bp) based on hypothesis test methods. In brief, the input genomic regions are broken down into overlapping fragments of equal length across the RRBS-selected regions. In each sliding window, regions that satisfy the following criteria are selected for further statistical testing: (1) each site in different samples of aligned reads meets the user-defined coverage threshold (default 4); (2) the number of selected type of cytosines (C/CG/CHG/CHH) should be larger than the user-defined value (default 5); and (3) the fold difference of mean methylation level (the maximum/minimum among samples for each region) should be larger than the user-defined value (default 1.5). Fold difference is calculated from expression (1) and (2) as follows:

$$\text{Methylation level} = \frac{\text{mC}(\#)}{\text{mC}(\#) + \text{umC}(\#)} \quad (1)$$

$$\text{Fold difference} = \frac{\text{maximum methylation level}}{\text{minimum methylation level}} \quad (2)$$

where mC (#) and umC (#) represents total number of methylated and unmethylated cytosines, respectively, from clean reads in given sliding window. Methylation level of corresponding sliding window region in each sample is calculated through expression (1). The maximum and minimum methylation level is then determined in sliding window from multiple samples, respectively. Last, the fold difference is calculated by expression (2).

For statistical testing of two or more samples, several statistical methods (Table 1), including both parametric modules and nonparametric modules, can be selected to identify DMRs. After the hypothesis tests are carried out, regions with P values less than the cut-off value (default 0.01) can be defined as putative DMRs. To control the FDR, P values of putative DMRs are corrected using the method proposed by Benjamini and Hochberg (1995) to filter out those regions with FDR values larger than the cut-off value (default 0.01). To extend adjacent candidate DMRs, regions closer than selected length (default 100 bp) are merged. Once the DMRs are identified, BEDTools is implemented for flexible annotation of DMRs by comparing the chromosome coordinates of DMRs with the corresponding annotation information in GFF/GTF/BED format [Quinlan and Hall, 2010].

Table 1. The Statistics Approaches Implemented in DMR Detection

Statistics method	Statistic model	Number of samples
T test	Parametric	Two
Wilcoxon test	Nonparametric	Two
Chi-square test	Nonparametric	Two
Fisher test	Nonparametric	Two
ANOVA	Parametric	Three or more
Kruskal–Wallis test	Nonparametric	Three or more

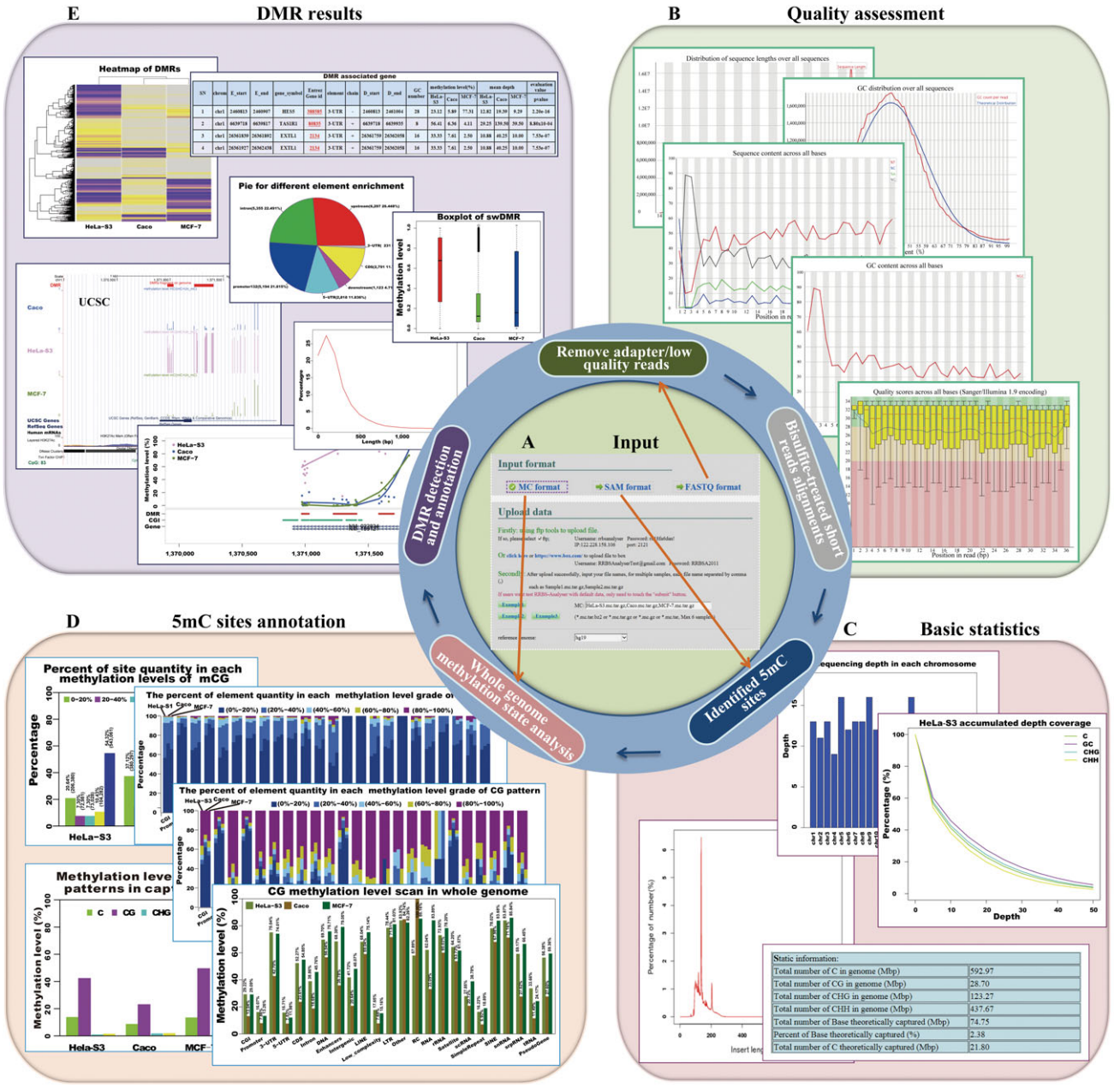


Figure 1. Snapshot of the results of RRBS-Analyser. **A:** FASTQ, SAM, or MC files are loaded as input along with several user-selected options. **B:** Quality assessment of raw data. **C:** Basic statistical information containing alignments, methylation, and bisulfite conversion rate. **D:** Detailed 5mC annotation information on different genomic features. **E:** The resulting DMR output, including DMR-associated genes, annotation, and visualization of DMR data in the genome browser or in the IGV programme with the "wig" format.

Data Input

RRBS-Analyser provides a simple and intuitive interface to allow users to flexibly analyze the DNA methylome generated from high-throughput sequencing (Fig. 1). The input requirement of RRBS-Analyser can be: (1) raw FASTQ reads, either single-end or paired-end produced by Illumina Solexa; (2) an alignment result in the generic SAM format [Li et al., 2009]; or (3) a MC file containing individual 5mC sites in the self-defined format, which considerably reduces the input size. To further reduce the input size, all input files can be compressed into .tar, .tar.gz, .gz, or .tar.bz2 formats. Notably, we found that the input size can sometimes be larger than 100 Mb for a MC file obtained from RRBS of human samples. Therefore, RRBS-Analyser has implemented the Box (https://www.box.com/), which is a cloud-based file storage and sharing service operating through a Web service application. In addition, users can also upload the data via our File Transfer Protocol (FTP) server, a link for which is available on the RRBS-Analyser Web server.

After successfully uploading the data to the Box or by FTP, users need to input the corresponding file name as a unique identifier of an uploaded file (for multiple samples, each file name must be separated by a comma).

Data Output

The RRBS-Analyser results can be retrieved by an assigned job ID, which is generated immediately after the data are uploaded successfully. A typical output contains four sections: quality assessment, basic statistics, 5mC sites annotation, and DMR result (Fig. 1). These sections are well organized with examples to help users with the correct input and to demonstrate the expected results.

The first section gives an overview of the raw reads quality, including CG dinucleotide distribution for all reads, practical and theoretical distribution of reads with varied GC contents, length distribution of clean reads filtered for adapters/low-quality bases, and base quality distribution.

The basic statistics section contains three parts: (1) global alignments information, which shows the number of raw reads and clean reads, the proportion aligned, the proportion uniquely aligned, and the number of mismatched raw reads; (2) methylation information, including overall methylation level (as CG/CHG/CHH), the number of 5mCs, and cytosines that match the reference genome as CG/CHG/CHH; and (3) the proportion of bisulfite conversion to indicate the conversion efficiency of bisulfite-treated DNA.

The annotation of 5mC sites section displays the 5mC sites list and detailed annotation information of methylated sites. First, RRBS-Analyser categorizes cytosines into CG, CHG, and CHH content to show their mean methylation levels. To further describe the quantity distribution of cytosines that are methylated or matched to the reference genome, RRBS-Analyser divides the methylation level into five scales: 0%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100%. Meanwhile, the methylation status in distinct genomic features is described as follows: quantity distribution of different methylation levels in functional elements and methylation levels in different functional elements.

The detailed DMR annotation information will be generated if users provide multiple samples. In this section, RRBS-Analyser shows DMR length distribution, a boxplot based on the methylation levels of DMRs (showing the DMR distribution at different methylation levels), and the DMR distribution in different functional elements and DMR-associated genes. RRBS-Analyser also provides the DMR coordinates, its associated gene region, mean methylation level, and sequencing depth, as well as *P* values and *q* values between different samples. To show the DMR cluster information, heatmap.2 in the R package is implemented to perform linkage hierarchical clustering of the methylation level for each DMR. RRBS-Analyser also provides the DMR information in the "wig" format, which can be used to display continuous DMR data in the UCSC genome browser or the integrative genomics viewer (IGV) program.

Implementation

RRBS-Analyser is constructed under an Apache/PHP/MySQL environment on the Red Hat Enterprise 5.5 Linux operating system. The back-end pipeline is implemented in Perl and R languages (http://www.r-project.org), which is run in parallel to accelerate the analysis process. All the plots are generated by R plot packages. The uploaded data will be analyzed on our high-performance computer with five computational nodes, each node containing four Quad-Core AMD processors (2.2 GHz each) and 32 GB of RAM. Meanwhile, RRBS-Analyser has a queuing module to control user-submitted jobs, which executes two jobs in parallel, with the remainder being put into a queue. When the submission is finished, the server will provide users with a job ID number, which can be used to retrieve the results once the job is finished or to reanalyze the data submitted previously. The Web client of RRBS-Analyser is implemented independently of operating systems and has been successfully tested with Microsoft Internet Explorer 8.0, Firefox 2/3, Google Chrome 24.0, and Safari 6.02 (under different versions of Linux, Microsoft Windows, and MacOS).

Perspectives

Rapid advances in NGS technologies have greatly facilitated genome-wide DNA methylome research [Meaburn and Schulz, 2012]. RRBS combines DNA digestion and size selection to perform high-throughput sequencing of a reproducible subset of the genome. It is indicated to be accurate and cost-efficient for DNA methylation studies at single-base resolution [Wang et al., 2012]. However, the vast amount of data generated by NGS poses multiple challenges for efficient data processing. At present, although many tools have been developed to process and analyze DNA methylation sequencing data, there are still no public online services available for comprehensive analysis of RRBS data. In addition, those tools that are available often need complex installation, redundant operations, and high-performance computational capability, and are not user-friendly for nonbioinformaticians.

Therefore, we have developed a novel and comprehensive platform, RRBS-Analyser, for the analysis of whole-genome shotgun RRBS data, which allows quality assessment of bisulfite-treated short reads, and detects and annotates 5mCs, as well as detecting and annotating DMRs based on multiple samples. We found that uploading large-size data is technically difficult during the development of RRBS-Analyser. Box cloud storage technology, which is an online file sharing and cloud content management service, is excellent for large-size data transmission. Thus, in RRBS-Analyser, Box storage technology is implemented for users to upload their RRBS sequencing data conveniently. In addition, FTP transmission is also supported by RRBS-Analyser. RRBS-Analyser is freely available for noncommercial use and will be updated regularly to keep up with the latest resources of the implemented databases. Currently, RRBS-Analyser can only analyze RRBS data based on the methylation insensitive enzyme MspI; more enzymes will be supported in future updates of the platform. As the Internet is constantly changing, a public interface that will be more suitable to cloud-based systems may emerge in the future; RRBS server will be updated and support the interface for the users. Additionally, RRBS-Analyser only supports nine reference genomes; more reference genomes will be added in the future.

Conclusion

We believe that RRBS-Analyser provides the scientific community with an integrated infrastructure for genome-wide investigation of DNA methylation, based on the large amount of data generated by RRBS, and it will be very useful for studies in the field of epigenomics.

References

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13:R87.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery rate: a practical and powerful approach to multiple testing. JSTOR 57:289–300.

Benoukraf T, Wongphayak S, Hadi LH, Wu M, Soong R. 2013. GBSA: a comprehensive software for analysing whole genome bisulfite sequencing data. Nucleic Acids Res 41:e55.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16:6–21.

Chen PY, Cokus SJ, Pellegrini M. 2010. BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics 11:203.

Coarfa C, Yu F, Miller CA, Chen Z, Harris RA, Milosavljevic A. 2010. Pash 3.0: a versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. BMC Bioinformatics 11:572.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci USA 89:1827–1831.

Grunau C, Schattevoy R, Mache N, Rosenthal A. 2000. MethTools—a toolbox to visualize and analyze DNA methylation data. Nucleic Acids Res 28:1053–1058.

Halachev K, Bast H, Albrecht F, Lengauer T, Bock C. 2012. EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome Biol 13:R96.

Hansen KD, Langmead B, Irizarry RA. 2012. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 13:R83.

Harris EY, Ponts N, Le Roch KG, Lonardi S. 2012. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. Bioinformatics 28:1795–1796.

Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, Olshen A, Ballinger T, et al. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol 28:1097–1105.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27:1571–1572.

Ku CS, Naidoo N, Wu M, Soong R. 2011. Studying the epigenome using next generation sequencing. J Med Genet 48:721–730.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, Zhang H, Chen Q, et al. 2010. The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol 8:e1000533.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322.

Meaburn E, Schulz R. 2012. Next generation sequencing in epigenetics: insights and challenges. Semin Cell Dev Biol 23:192–199.

Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res 33:5868–5877.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Smith AD, Chung WY, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ. 2009. Updates to the RMAP short-read mapping software. Bioinformatics 25:2841–2842.

Sun Z, Baheti S, Middha S, Kanwar R, Zhang Y, Li X, Beutler AS, Klee E, Asmann YW, Thompson EA, Kocher JP. 2012. SAAP-RRBS: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing. Bioinformatics 28:2180–2181.

Wang L, Sun J, Wu H, Liu S, Wang J, Wu B, Huang S, Li N, Zhang X. 2012. Systematic assessment of reduced representation bisulfite sequencing to human blood samples: a promising method for large-sample-scale epigenomic studies. J Biotechnol 157:1–6.

Warnecke PM, Stirzaker C, Song J, Grunau C, Melki JR, Clark SJ. 2002. Identification and resolution of artifacts in bisulfite sequencing. Methods 27:101–107.

Xi Y, Bock C, Muller F, Sun D, Meissner A, Li W. 2012. RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. Bioinformatics 28:430–432.

Xin Y, Ge Y, Haghighi FG. 2011. Methyl-Analyzer—whole genome DNA methylation profiling. Bioinformatics 27:2296–2297.

Zhang Y, Liu H, Lv J, Xiao X, Zhu J, Liu X, Su J, Li X, Wu Q, Wang F, Cui Y. 2011. QDMR: a quantitative method for identification of differentially methylated regions by entropy. Nucleic Acids Res 39:e58.

RESEARCH PAPER

RNA Biology 10:6, 1–6; June 2013; © 2013 Landes Bioscience

mirTools 2.0 for non-coding RNA discovery, profiling and functional annotation based on high-throughput sequencing

Jinyu Wu,[1,2] Qi Liu,[2] Xin Wang,[2] Jiayong Zheng,[3] Tao Wang,[2] Mingcong You,[2] Zhong Sheng Sun[2,4,*] and Qinghua Shi[1,*]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences; University of Science and Technology of China; Hefei, China; [2]Institute of Genomic Medicine; Wenzhou Medical College; Wenzhou, China; [3]Department of Laboratory Medicine; Third People's Hospital of Wenzhou; Wenzhou, China; [4]Beijing Institutes of Life Science; Chinese Academy of Sciences; Beijing, China

Keywords: ncRNA, miRNA, miRNA targets, web server, mirTools, next-generation sequencing

Next-generation sequencing has been widely applied to understand the complexity of non-coding RNAs (ncRNAs) in a cost-effective way. In this study, we developed mirTools 2.0, an updated version of mirTools 1.0, which includes the following new features. (1) From miRNA discovery in mirTools 1.0, mirTools 2.0 allows users to detect and profile various types of ncRNAs, such as miRNA, tRNA, snRNA, snoRNA, rRNA and piRNA. (2) From miRNA profiling in mirTools 1.0, mirTools 2.0 allows users to identify miRNA-targeted genes and performs detailed functional annotation of miRNA targets, including Gene Ontology, KEGG pathway and protein-protein interaction. (3) From comparison of two samples for differentially expressed miRNAs in mirTools 1.0, mirTools 2.0 allows users to detect differentially expressed ncRNAs between two experimental groups or among multiple samples. (4) Other significant improvements include strategies used to detect novel miRNAs and piRNAs, more taxonomy categories to discover more known miRNAs and a stand-alone version of mirTools 2.0. In conclusion, we believe that mirTools 2.0 (122.228.158.106/mr2_dev and centre.bioinformatics.zj.cn/mr2_dev) will provide researchers with more detailed insight into small RNA transcriptomes.

Introduction

Non-coding RNA (ncRNAs) has been increasingly recognized as an important molecular in the past few years.[1] Among them, microRNA (miRNA) is small, approximately 19–25 nt RNA molecule, which is involved in post-transcriptional regulation of gene expression. It plays important roles in regulation of numerous biological processes, such as development, cell differentiation and proliferation, apoptosis and metabolism.[2] Small nuclear RNA (snRNA) is primarily involved in RNA splicing and assists in the regulation of transcription factors and maintains telomeres.[3] Small nucleolar RNA (snoRNA) plays a crucial role in modification of target RNAs and processing of rRNA during ribosome subunit synthesis.[4] Piwi-interacting RNA (piRNA), an approximately 24–31 nt RNA molecule, plays an important role in regulation of cell division and maintenance of germline stem cells.[5] A recent study showed that piRNA is also involved in epigenetic control of memory-related synaptic plasticity in neural cells.[6]

Next-generation sequencing (NGS) has been widely applied to characterize small RNA transcriptomes under various conditions. It provides an unprecedented opportunity to discover ncRNAs and identify differentially expressed ncRNA transcripts.[7] However, the massive amount of data generated by NGS poses great bioinformatics challenges for detection and functional annotation of ncRNAs. Therefore, a number of computational methods have been developed for mining small RNA sequencing data. Among them, many tools have mainly focused on miRNA analysis, such as miRDeep,[8-10] Mireval,[11] miRFinder,[12] miRNAkey,[13] miRanalyzer,[14] miRExpress,[15] miRTRAP,[16] DSAP[17] and MIReNA.[18] In addition, several integrated ncRNA analysis tools have been released, such as SeqCluster,[19] DARIO,[20] ncPRO-seq,[21] CPSS,[22] Shortran,[23] NORAHDESK,[24] APART[25] and smyRNA.[26]

We previously developed a web service, mirTools 1.0, which provides annotation of miRNAs based on NGS and has been widely used.[27] Comprehensive comparison and evaluation of bioinformatics tools for miRNA deep-sequencing has indicated that mirTools 1.0 has a good performance for miRNA coverage, accuracy and sensitivity, as well as computational time.[28] However, in the past 2 y, we have received considerable feedback from users. These users expected us to update mirTools to include more versatile functions, such as miRNA-targeted genes and further functional annotation, other types of ncRNAs besides miRNAs and multiple sample comparison. Therefore, in this study, an integrated web server, mirTools 2.0, an updated version of mirTools 1.0, was developed to investigate ncRNA sequences, expression levels, differentially expressed ncRNAs and miRNA-targeted

*Correspondence to: Zhong Sheng Sun; Email: sunzs@mail.biols.ac.cn; Qinghua Shi; Email: qshi@ustc.edu.cn
Submitted: 02/24/13; Revised: 05/19/13; Accepted: 05/28/13
http://dx.doi.org/10.4161/rna.25193

genes and their functional annotation, which will be valuable for deciphering the functional roles of ncRNAs hidden in the large amount of NGS data.

Results and Discussion

Implementation. The web server mirTools 2.0 is constructed under the Apache/PHP/MySQL environment in the Linux system. The back-end pipeline is implemented in Perl language and the plots are generated by R packages (www.r-project.org). Compared with mirTools 1.0, the computational power of mirTools 2.0 has been enhanced and it is equipped with four Quad-Core AMD processors (2.2 GHz each) and 32 GB of RAM. It will only take approximately 30 min to detect and quantify ncRNAs for a given sample (~10 Mb size). Additionally, the queuing module can execute more jobs in parallel.

Data input. The web server mirTools 2.0 provides more functional modules than mirTools 1.0, including a single case, two cases, group cases and re-analysis. The single case module allows users to detect various types of known and novel ncRNAs, and performs functional annotation of the miRNA-targeted genes for a single sample. Two cases and group cases modules allow users to identify differentially expressed ncRNAs between or among samples. The re-analysis module is designed to allow users to run previously submitted data with adjustable parameters, which avoids resubmitting the sample data.

In single case and two case modules, similar to mirTools 1.0, the input of mirTools 2.0 is a trimmed FASTA file where all the identical raw reads are aggregated and cleaned into a non-redundant FASTA file to reduce the input size. To further reduce the input size, the FASTA file can be compressed in rar, zip or gz formats, with a maximum size of 30 Mb. In addition, mirTools 2.0 supports the input of original mapped reads in SAM/BAM format, which can be generated by many public alignment software, such as Bowtie (bowtie-bio.sourceforge.net) and BWA (bio-bwa.sourceforge.net). In the group case module, the expression table files of ncRNAs are required, which can be obtained from the single case and two case modules. Users can directly input a single case analysis job ID and the web server will retrieve the corresponding expression table file automatically. In all modules, the mail address is optional and the web server will give users a job ID, which can be used to retrieve the results once the job is finished or to reanalyze the data submitted previously.

Data output. The mirTool 2.0 results are presented in intuitive HTML pages, of which a typical output consists of six parts: basic statistics, annotation, miRNA, miRNA targets, ncRNA and differential expression (**Fig. 1**). The basic statistics output contains length distribution charts of short reads, pie charts summary of reference genome mapping and the chromosome distribution. The annotation output includes pie charts of mapped reads with different functional categories, ncRNA distribution and repeat-associated RNA distribution. The web server mirTools 2.0 plots the unique read distribution and its expression levels (the number of reads for each tag reflects its relative abundance).

The miRNA output consists of known miRNAs and putative novel miRNAs. The detailed annotation of each miRNA contains the miRNA name, arm on the hairpin, absolute count, relative count, pre-miRNA number (hairpin secondary structure for novel miRNAs) and related expression information of the most abundant tag. In addition, users can view the read mapping information and miRNA isoforms in a pop-up webpage by clicking the "pre-miRNA" link.

The miRNA targets output contains the predicted miRNA-targeted genes and their functional annotation with GO, the KEGG pathway and the PPI network. The miRNA-targeted genes tables display the miRNA name, the targeted gene name, the minimum free energy, the score value (P value for RNAhybrid) and target prediction tool used. The known miRNA-targeted genes tables also contain the "other tools" column to indicate whether the targets are supported by other tools. The GO and KEGG pathway annotation tables illustrate the enriched GO terms and pathway terms of targeted genes predicted, respectively, which can be sorted by enrichment fold and P value. The PPI annotation tables depict the protein interaction information of miRNA targets in STRING databases. Users can visualize the interaction intuitively in the implemented Cytoscape Web, which supports node dragging and searching, by clicking the "show the network in Cytoscape" link.

The ncRNA output shows information of other ncRNAs, except for miRNAs and their expression level. Information on the identified known piRNAs and novel piRNAs are also included in this output. The detailed annotation of each ncRNA contains the ncRNA name, absolute count, relative count, hairpin number and related expression information of the most abundant tag.

In a two-sample study, the differential expression output contains expression correlation dot charts and differentially expressed ncRNAs lists between the two samples. The annotation information of the differential expression list contains the ncRNA name, sample "a" relative count, sample "b" relative count, the fold change, the up/down tag and the P value. In group case results, differentially expressed ncRNAs between two groups are listed. The annotation information of the group expression list also contains the expression value of each sample, the median expression value of the group, the up/down tag and the statistical P value. All these components are well organized with examples to facilitate users with correct input and expected results.

Discussion

NGS has greatly facilitated RNA transcriptome studies, among which small RNA sequencing offers a cost-effective and in-depth method to comprehensively investigate ncRNAs in a genome-wide manner.[7] However, one of the main challenges lies with the analysis of miRNAs and other ncRNAs from the large amount of sequencing data. The web server mirTools 2.0 was developed for research communities toward a fully automated and easy to use web service suitable for ncRNA discovery, profiling and functional annotation based on high-throughput sequencing.

The web server mirTools 2.0 is freely available for non-commercial use and will be updated regularly to keep up with the latest annotation information of the implemented databases. In mirTools 1.0, we received a lot of valuable feedback and

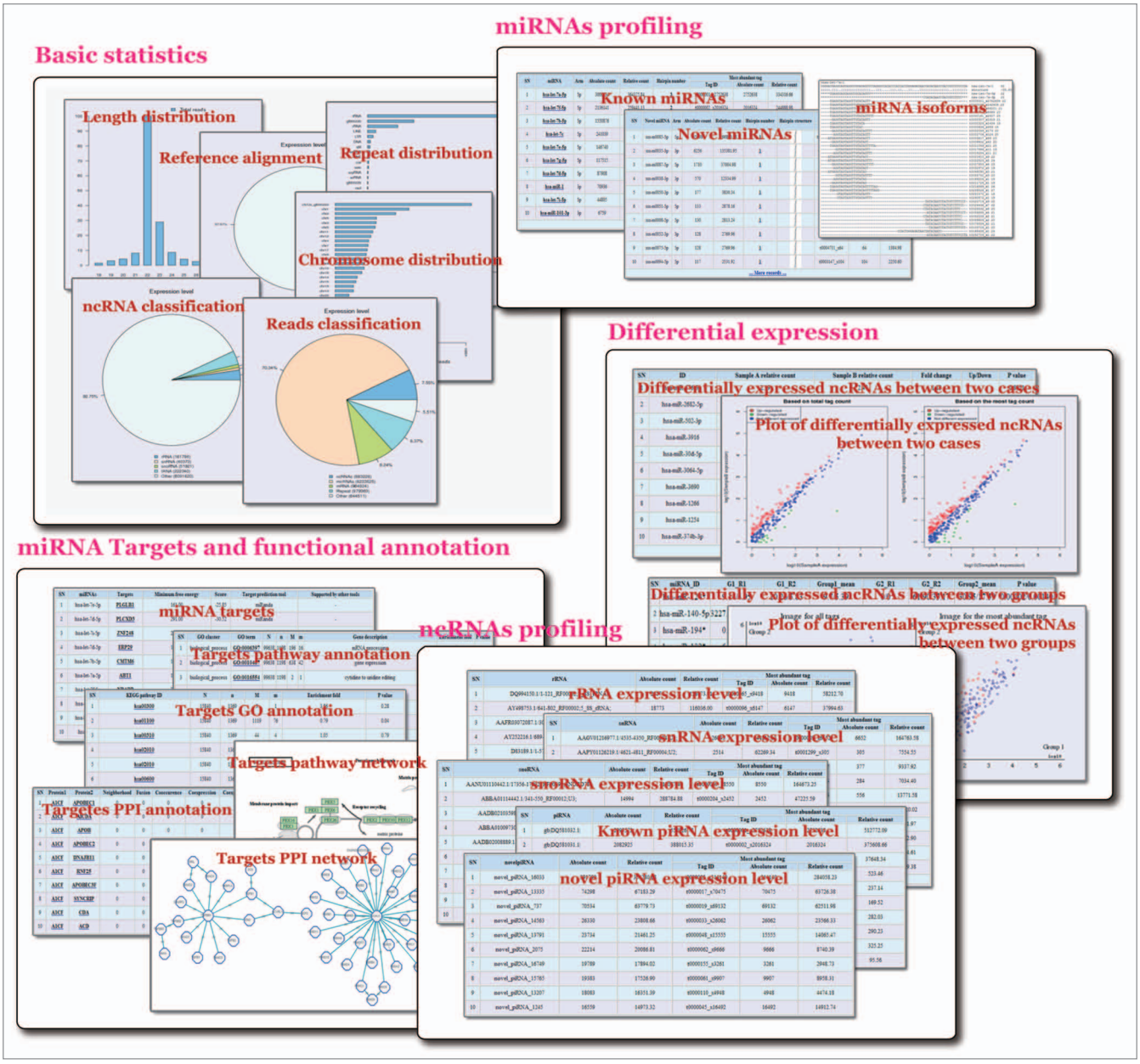


Figure 1. Output screenshots of mirTools 2.0. The output includes: (1) basic statistics, such as length distribution, percentage of reads aligned to the reference genome, chromosome distribution, functional categories of reads and ncRNA distribution; (2) known miRNA, putative novel miRNA, miRNA isoforms and modification; (3) miRNA-targeted genes and functional annotation based on GO, the KEGG pathway and the PPI network; (4) expression information of other types of ncRNA, such as rRNA, tRNA, snRNA, snoRNA and piRNA; and (5) differentially expressed ncRNAs between two cases, two experimental groups or among multiple samples.

suggestions from users worldwide, and this feedback has been helpful for developing mirTools 2.0. Therefore, we sincerely welcome questions, comments and suggestions, which will be useful for feedback for the enhanced function of mirTools 2.0. Currently, mirTools 2.0 can only detect the known ncRNAs, novel miRNAs and novel piRNAs. In the future, we will develop or incorporate a tool to predict other type of novel ncRNAs. In the meantime, phylogenetic conservation analysis of ncRNAs across different species will be provided. Considering the network limits, currently, the maximum file upload size is 30 Mb, regardless of whether compression is involved. Therefore, we have developed a stand-alone version of mirTools 2.0 to allow users to run it on their own server. In the future, we will design an FTP module to allow users to submit larger data to enhance the usability of web server. In conclusion, we believe that mirTools 2.0 will provide the scientific community with an integrated web server to assist research for identifying various types of ncRNA, profiling expression levels, predicting miRNA targeted genes and functional annotation based on the large amount of data generated from NGS.

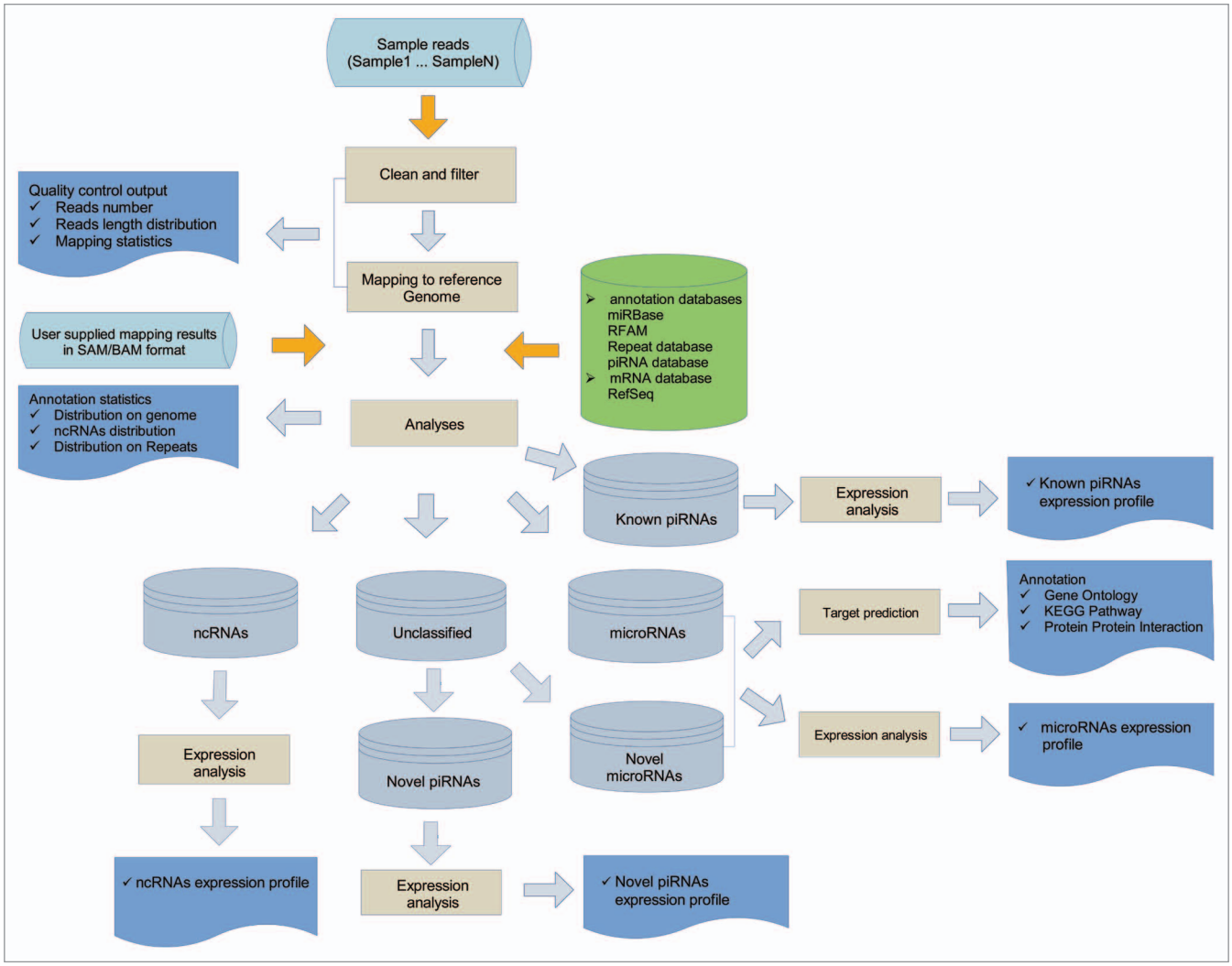


Figure 2. The overall workflow of mirTools 2.0. The workflow includes clean and filtering raw reads, alignment of them to the reference genome, classification of aligned reads, detection of expression levels of various types of ncRNAs and the differentially expressed ncRNAs between two cases/two experiment groups or among multiple samples, prediction of novel miRNAs and piRNAs, identification of miRNA targeted genes and further functional annotation based on GO, the KEGG pathway and the PPI network.

Materials and Methods

Overview of the workflow of mirTools 2.0. The overall workflow of mirTools 2.0 is shown in **Figure 2**. Briefly, mirTools 2.0 filters out raw reads to exclude low quality and 3/5' adaptor sequences and trim them into clean reads. Clean reads are then mapped onto the reference genome and mapping results are converted into the SAM/BAM format with SAMtools (samtools.sourceforge.net) to serve as a generic alignment format compatible with different alignment tools. Based on public resources, the mapped reads are annotated and classified into known ncRNAs. Novel miRNAs and piRNAs will be predicted from unclassified aligned reads. miRNA-targeted genes and further functional annotations are conducted for both known and novel miRNAs based on a number of implemented tools. Finally, all the results are shown in different types of tables and figures on an HTML page, and these are available for downloading intermediate annotation results and the final results.

Discovery and profiling of known and novel ncRNAs. To identify known and novel ncRNAs, sequence reads are first aligned to the reference genome using the SOAP program.[29] Subsequently, aligned reads are associated with the annotation information of several public databases. In addition to miRBase (www.mirbase.org), Rfam (rfam.sanger.ac.uk), repeat database produced by RepeatMasker (www.repeatmasker.org) and coding genes of the reference genome, piRNA from the piRNABank database (pirnabank.ibab.ac.in) is also incorporated to identify known piRNAs. Currently, mirTools 2.0 is compatible for use with 32 reference genomes across vertebrates, insects, deuterostomes, nematodes and plants. The aligned reads are classified into known miRNAs, other types of ncRNAs, known piRNAs, repeat-associated RNA and mRNAs. miRNA isoforms and modification can be obtained through changing the mismatch

number in the SOAP program. The ncRNAs reads annotated by Rfam are further classified into sRNA, tRNA, snRNAs and snoRNA.

The unclassified aligned reads termed as "unclassified" are used to detect novel miRNAs and piRNAs. In mirTools 1.0, we used the miRDeep program to predict novel miRNAs. In mirTools 2.0, we implemented a new version of miRDeep[9] to discover novel miRNAs. We also implemented another broadly used program, Mireap (sourceforge.net/projects/mireap), which combines secondary structure, minimum free energy, Dicer cleavage site, small RNA position and depth, to discover novel miRNAs from NGS.[28,30] During this process, the secondary structures are predicted using the RNAfold program in Vienna RNA package (www.tbi.univie.ac.at/RNA). The remaining unclassified reads are used to detect novel piRNAs using a k-mer scheme, which has been indicated to be high accuracy and specificity for predicting novel piRNAs.[31]

Identification of miRNA-targeted genes and functional annotation. To identify known and novel miRNA targeted genes, mirTools 2.0 implements two widely used tools miRanda (www.microrna.org) and RNAhybrid (bibiserv.techfak.uni-bielefeld.de/rnahybrid/). In addition, miRNA-targeted gene results from another six tools or databases are also integrated, including TargetScan (www.targetscan.org), TargetSpy (www.targetscan.org), miRNAMap (mirnamap.mbc.nctu.edu.tw), microT v4.0 (diana.cslab.ece.ntua.gr/microT/), MicroCosm (www.ebi.ac.uk/enright-srv/microcosm) and MirTarget2 (mirdb.org).

To explore the potential biological function of predicted miRNA-targeted genes, we annotated them with Gene Ontology (GO), the KEGG pathway and the protein-protein interaction (PPI) network. For GO analysis, the predicted targets are mapped to the GO annotation data set to extract their GO annotation,[32] and then Fisher's exact test is used to perform GO enrichment analysis (enrichment ratio > 2 and P value < 0.01 at default). Pathway assignment information of miRNA-targeted genes is extracted from the KEGG pathway database[33] and corresponding enrichment analysis is performed using the hypergeometric test (enrichment ratio > 2 and P value < 0.01 at default). Moreover, PPI annotation of miRNA-targeted genes is retrieved from the STRING database.[34] Visualization of the PPI network can be conducted using the implemented Cytoscape Web tool, which is an interactive web-based network browser that allows easy displaying of graphs.[35]

Detection of differentially expressed ncRNAs. To determine the relative ncRNA expression level and its abundance, each identified ncRNA read count is normalized to the total read count of its belonging type of ncRNA to obtain reads per million (RPM) value. Similar to mirTools 1.0, mirTools 2.0 has two strategies to estimate the expression level of a given ncRNA: the relative total read count and the most abundant read (often considered as mature miRNA). To detect differentially expressed ncRNAs between two samples, the Bayesian method is used to calculate the statistical significance (P value) based on the relative total read count and most abundant read count.[36]

In addition, we developed a group case module, which can compare the difference within and between experimental groups with multiple replicates or samples. If a specific experimental group has two conditions, the Wilcoxon Rank-sum test is applied to infer the statistical significant difference. If a specific experimental group has more than two conditions, the Kruskal-Wallis H test is applied to infer the statistical significant difference. The Wilcoxon Rank-sum test is used to identify the differentially expressed ncRNAs between experimental groups. In all conditions, at default, a specific ncRNA is considered to be differentially expressed with a P value < 0.01 and a fold change > 2.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grant no. 31171236), the National High Technology Research and Development Program of China (2012AA02A201), the Key Science and Technology Innovation Team of Zhejiang Province (2012R10048-05) and the International S&T Cooperation Program of China (2011DFA30670).

References

1. David R. Non-coding RNAS: A new member of the family. Nat Rev Mol Cell Biol 2012; 13:686; PMID:23011341; http://dx.doi.org/10.1038/nrm3449
2. Ebert MS, Sharp PA. Roles for microRNAs in conferring robustness to biological processes. Cell 2012; 149:515-24; PMID:22541426; http://dx.doi.org/10.1016/j.cell.2012.04.005
3. Karijolich J, Yu YT. Spliceosomal snRNA modifications and their function. RNA Biol 2010; 7:192-204; PMID:20215871; http://dx.doi.org/10.4161/rna.7.2.11207
4. Taft RJ, Glazov EA, Lassmann T, Hayashizaki Y, Carninci P, Mattick JS. Small RNAs derived from snoRNAs. RNA 2009; 15:1233-40; PMID:19474147; http://dx.doi.org/10.1261/rna.1528909
5. Juliano C, Wang J, Lin H. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. Annu Rev Genet 2011; 45:447-69; PMID:21942366; http://dx.doi.org/10.1146/annurev-genet-110410-132541
6. Rajasethupathy P, Antonov I, Sheridan R, Frey S, Sander C, Tuschl T, et al. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. Cell 2012; 149:693-707; PMID:22541438; http://dx.doi.org/10.1016/j.cell.2012.02.057
7. Zhou L, Li X, Liu Q, Zhao F, Wu J. Small RNA transcriptome investigation based on next-generation sequencing technology. J Genet Genomics 2011; 38:505-13; PMID:22133681; http://dx.doi.org/10.1016/j.jgg.2011.08.006
8. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 2008; 26:407-15; PMID:18392026; http://dx.doi.org/10.1038/nbt1394
9. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 2012; 40:37-52; PMID:21911355; http://dx.doi.org/10.1093/nar/gkr688
10. Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics 2011; 27:2614-5; PMID:21775303
11. Ritchie W, Théodule FX, Gautheret D. Mireval: a web tool for simple microRNA prediction in genome sequences. Bioinformatics 2008; 24:1394-6; PMID:18453555; http://dx.doi.org/10.1093/bioinformatics/btn137
12. Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, Zhao SH. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics 2007; 8:341; PMID:17868480; http://dx.doi.org/10.1186/1471-2105-8-341
13. Ronen R, Gan I, Modai S, Sukacheov A, Dror G, Halperin E, et al. miRNAkey: a software for microRNA deep sequencing analysis. Bioinformatics 2010; 26:2615-6; PMID:20801911; http://dx.doi.org/10.1093/bioinformatics/btq493

14. Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 2009; 37(Web Server issue):W68-76; PMID:19433510; http://dx.doi.org/10.1093/nar/gkp347
15. Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinformatics 2009; 10:328; PMID:19821977; http://dx.doi.org/10.1186/1471-2105-10-328
16. Hendrix D, Levine M, Shi W. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome Biol 2010; 11:R39; PMID:20370911; http://dx.doi.org/10.1186/gb-2010-11-4-r39
17. Huang PJ, Liu YC, Lee CC, Lin WC, Gan RR, Lyu PC, et al. DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res 2010; 38(Web Server issue):W385-91; PMID:20478825; http://dx.doi.org/10.1093/nar/gkq392
18. Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics 2010; 26:2226-34; PMID:20591903; http://dx.doi.org/10.1093/bioinformatics/btq329
19. Pantano L, Estivill X, Martí E. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. Bioinformatics 2011; 27:3202-3; PMID:21976421; http://dx.doi.org/10.1093/bioinformatics/btr527
20. Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 2011; 39(Web Server issue):W112-7; PMID:21622957; http://dx.doi.org/10.1093/nar/gkr357
21. Chen CJ, Servant N, Toedling J, Sarazin A, Marchais A, Duvernois-Berthet E, et al. ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. Bioinformatics 2012; 28:3147-9; PMID:23044543; http://dx.doi.org/10.1093/bioinformatics/bts587
22. Zhang Y, Xu B, Yang Y, Ban R, Zhang H, Jiang X, et al. CPSS: a computational platform for the analysis of small RNA deep sequencing data. Bioinformatics 2012; 28:1925-7; PMID:22576177; http://dx.doi.org/10.1093/bioinformatics/bts282
23. Gupta V, Markmann K, Pedersen CN, Stougaard J, Andersen SU. shortran: a pipeline for small RNA-seq data analysis. Bioinformatics 2012; 28:2698-700; PMID:22914220; http://dx.doi.org/10.1093/bioinformatics/bts496
24. Ragan C, Mowry BJ, Bauer DC. Hybridization-based reconstruction of small non-coding RNA transcripts from deep sequencing data. Nucleic Acids Res 2012; 40:7633-43; PMID:22705792; http://dx.doi.org/10.1093/nar/gks505
25. Zywicki M, Bakowska-Zywicka K, Polacek N. Revealing stable processing products from ribosome-associated small RNAs by deep-sequencing data analysis. Nucleic Acids Res 2012; 40:4013-24; PMID:22266655; http://dx.doi.org/10.1093/nar/gks020
26. Salari R, Aksay C, Karakoc E, Unrau PJ, Hajirasouliha I, Sahinalp SC. smyRNA: a novel Ab initio ncRNA gene finder. PLoS One 2009; 4:e5433; PMID:19415115; http://dx.doi.org/10.1371/journal.pone.0005433
27. Zhu E, Zhao F, Xu G, Hou H, Zhou L, Li X, et al. mirTools: microRNA profiling and discovery based on high-throughput sequencing. Nucleic Acids Res 2010; 38(Web Server issue):W392-7; PMID:20478827; http://dx.doi.org/10.1093/nar/gkq393
28. Li Y, Zhang Z, Liu F, Vongsangnak W, Jing Q, Shen B. Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. Nucleic Acids Res 2012; 40:4298-305; PMID:22287634; http://dx.doi.org/10.1093/nar/gks043
29. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009; 25:1966-7; PMID:19497933; http://dx.doi.org/10.1093/bioinformatics/btp336
30. Cheng WC, Chung IF, Huang TS, Chang ST, Sun HJ, Tsai CF, et al. YM500: a small RNA sequencing (smRNA-seq) database for microRNA research. Nucleic Acids Res 2013; 41(Database issue):D285-94; PMID:23203880; http://dx.doi.org/10.1093/nar/gks1238
31. Zhang Y, Wang X, Kang L. A k-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics 2011; 27:771-6; PMID:21224287; http://dx.doi.org/10.1093/bioinformatics/btr016
32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al.; The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet 2000; 25:25-9; PMID:10802651; http://dx.doi.org/10.1038/75556
33. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010; 38(Database issue):D355-60; PMID:19880382; http://dx.doi.org/10.1093/nar/gkp896
34. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009; 37(Database issue):D412-6; PMID:18940858; http://dx.doi.org/10.1093/nar/gkn760
35. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. Bioinformatics 2010; 26:2347-8; PMID:20656902; http://dx.doi.org/10.1093/bioinformatics/btq430
36. Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 1997; 7:986-95; PMID:9331369

JMG Online First, published on January 16, 2015 as 10.1136/jmedgenet-2014-102656

ORIGINAL ARTICLE

mirTrios: an integrated pipeline for detection of de novo and rare inherited mutations from trios-based next-generation sequencing

Jinchen Li,[1,2,3] Yi Jiang,[2] Tao Wang,[2] Huiqian Chen,[2] Qing Xie,[2] Qianzhi Shao,[2] Xia Ran,[2] Kun Xia,[3] Zhong Sheng Sun,[1,2] Jinyu Wu[1,2]

ABSTRACT

Objectives Recently, several studies documented that de novo mutations (DNMs) play important roles in the aetiology of sporadic diseases. Next-generation sequencing (NGS) enables variant calling at single-base resolution on a genome-wide scale. However, accurate identification of DNMs from NGS data still remains a major challenge. We developed mirTrios, a web server, to accurately detect DNMs and rare inherited mutations from NGS data in sporadic diseases.

Methods The expectation-maximisation (EM) model was adopted to accurately identify DNMs from variant call files of a trio generated by GATK (Genome Analysis Toolkit). The GATK results, which contain certain basic properties (such as PL, PRT and PART), are iteratively integrated into the EM model to strike a threshold for DNMs detection. Training sets of true and false positive DNMs in the EM model were built from whole genome sequencing data of 64 trios.

Results With our in-house whole exome sequencing datasets from 20 trios, mirTrios totally identified 27 DNMs in the coding region, 25 of which (92.6%) are validated as true positives. In addition, to facilitate the interpretation of diverse mutations, mirTrios can also be employed in the identification of rare inherited mutations. Embedded with abundant annotation of DNMs and rare inherited mutations, mirTrios also supports known diagnostic variants and causative gene identification, as well as the prioritisation of novel and promising candidate genes.

Conclusions mirTrios provides an intuitive interface for the general geneticist and clinician, and can be widely used for detection of DNMs and rare inherited mutations, and annotation in sporadic diseases. mirTrios is freely available at http://centre.bioinformatics.zj.cn/mirTrios/.

INTRODUCTION

De novo mutations (DNMs), arising from meiosis of the gametes of the parents (ie, sperm and egg) and transmitted to their child, usually have severe biological or phenotypic consequences when they affect functionally important nucleotides in the genome.[1] DNMs represent the most extreme form of rare genetic mutation and make these mutations prime candidates for causing sporadic genetic diseases that remain in a population despite the reduced fecundity.[2 3] The widespread availability of next-generation sequencing (NGS), such as whole exome sequencing (WES) and whole genome sequencing (WGS), revolutionised the identification of DNMs on a genome-wide scale. Attention has been mostly focused on neuropsychiatric diseases,[1–5] such as autism spectrum disorders (ASDs), schizophrenia, intellectual disability, and epileptic encephalopathy. These studies serve as pioneers, and many more large scale studies of other genetic diseases (such as congenital heart disease[6]) by NGS to identify risk-associated DNMs are underway.[5 7]

With the development of NGS, a number of computational methods that address multi-sample (typically parent–offspring trios) variant detection and genotype calling have been developed, such as SAMtools,[8] GATK (Genome Analysis Toolkit),[9] TrioCaller,[10] VarScan,[11] Famseq[12] and VariantMaster.[13] Among them, FamSeq builds on Bayesian networks to provide the probability for each genotype of each variant using data from all familial members. These methods greatly increase the power of inferring genotypes and haplotypes, but if we directly apply these methods for DNM calling, the false discovery rate will be above 60%.[14] The potential error during PCR, sequencing and mapping may contribute to the false positive rate. In some cases, assumed DNMs are actually inherited mutations due to the low evenness in local genomic regions of multiple samples. Subsequently, PolyMutt,[15] DeNovoGear[16] and DNMFilter[17] were specifically developed for DNM detection from trio-based NGS. PolyMutt and DeNovoGear investigate all available family members jointly based on likelihood framework and likelihood-based error modelling, respectively. Both algorithms relied on the average mutation rate of each class of mutations across the given genome, while de novo mutation rates were found to vary strikingly across different genomes and regions.[18] DNMFilter is based on a machine-learning filtering approach to identify DNMs, the efficacy of which is sensitive to the training set. Recently, Scalpel was specifically developed to detect de novo and transmitted insertions and deletions (indels) in exome-capture data on the basis of localised assembly.[19] However, all the above software require a certain level of computational skills that can handle installing, minor processing of input raw data or even debugging when incompatibility of datasets occurs. There are still no public user-friendly online services available for comprehensive analysis from family-based NGS data in sporadic diseases. In this study, mirTrios, a web server implementing the expectation–

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2014-102656).

[1]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China
[2]Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, China
[3]State Key Laboratory of Medical Genetics, Central South University, Changsha, China

Correspondence to
Professor Jinyu Wu, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China; wujy@mail.biols.ac.cn
Professor Zhong Sheng Sun, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China; sunzs@mail.biols.ac.cn

JL and YJ contributed equally.

Received 17 July 2014
Accepted 16 December 2014

To cite: Li J, Jiang Y, Wang T, *et al*. *J Med Genet* Published Online First: [*please include* Day Month Year] doi:10.1136/jmedgenet-2014-102656

maximisation (EM) algorithm, was developed to accurately identify DNMs from trio-based or family-based variant call file (VCF) results from NGS in sporadic diseases.

Studies have revealed that rare inherited variants, existing in homozygous, hemizygous, compound heterozygous, or dominant heterozygous forms, also make substantial contributions to sporadic diseases.[20–23] Thus, the identification of rare inherited mutations and the annotation of them are also provided in mirTrios. More importantly, the application of available online tools for identification of candidate genes in sporadic diseases is still insufficient. For analysis of multiple families, an adjusted TADA (Transmission And De novo Association) model[24] was used to prioritise candidate genes and provide a p value for statistical evidence in sporadic genetic diseases on the basis of extensive annotation.

METHODS

Accurate model for identification of DNMs

A generic VCF format file generated by GATK containing variant information of trios is required for the detection of DNMs. Due to the errors that occurred during sequencing, mapping and the variant calling process, discovering them simply by filtering based on the allowed scope of parameters, such as depth, quality and genotype, may not be sufficient to downsize false positive variants. Therefore, an EM algorithm adopted by mirTrios is used to further extract potential DNMs with closely related properties available from the VCF file (figure 1). These properties were iteratively integrated into the EM algorithm to strike a threshold for the identification of DNMs. The EM algorithm encompasses two major iterative steps:

(1) Expectation step (E step), calculating log-likelihood function on the basis of initial parameters or iterative results yielded in previous steps (the initial values were determined on the basis of a large amount of training data):

$$P(X,Z|\theta)=\sum_{i=1}^{n}\log p(x_i,z_i|\theta)=\sum_{i=1}^{n}\log\left(\pi_i N\left(x_i;\mu_{z_i},\sum_{z_i}\right)\right)$$

In this formula, $P(X,Z|\theta)$ represents the log-likelihood of variable X in the Gaussian mixture distribution Z with different iterative process θ. In addition, n denotes the total number of Gaussian distribution, and π_i denotes the weight of Gaussian distribution N in the iterative progresses. Every Gaussian mixture distribution z_i has a variable of x_i, a mean value of μ_{z_i} and a variable of Σ_{z_i}.

Expectation of the conditional distribution $p(X,Z|\theta^{old})$:

$$Q(\theta,\theta^{old})=E[\log p(X,Z|\theta),\theta^{old}]$$

(2) Maximisation step (M step): new parameters are generated by maximising the log-likelihood function, replacing θ^{old} with θ^{new} to obtain a maximised expectation $Q(\theta,\theta^{old})$. In these steps, θ^{old} represents the previous iterative process, andθ^{new} represents the current iterative process. Z represents Gaussian mixture distribution, and $\mu_{z_i}$and Σ_{z_i} represents the mean value and square deviation, respectively.

Generally, both the number of DNMs and non-DNMs from the large amount of trio samples present normal distributions (Gaussian distribution, Kolmogorov-Smirnov test, $p<0.001$), and jointly demonstrates a Gaussian mixture distribution. Based on the sample of Gaussian mixture distribution, we adopted the above described EM model to distinguish DNMs and non-DNMs, resulting in the probability:

$$P(x)=\sum_{i=1}^{n}\pi_i N\left(x_i;\mu_{z_i},\sum_{z_i}\right)$$

In the formula, $\sum_{i=1}^{N}\pi_i=1$, and $0\le\pi_i\le 1$; n denotes the total number of normal distribution; π_i denotes the weighting coefficient of each normal distribution represented by $N(x_i;\mu_{z_i},\Sigma_{z_i})$. The variable x_i is distributed normally with a mean value of μ_{z_i} and a variance of Σ_{z_i}.

All the DNM-related properties from VCF results generated by GATK are integrated into an EM model, which will be applied iteratively to strike a threshold for each variable that is essential for detection of DNMs. Several properties, QUAL (quality of alignment), Depth (total sequencing depth), QD

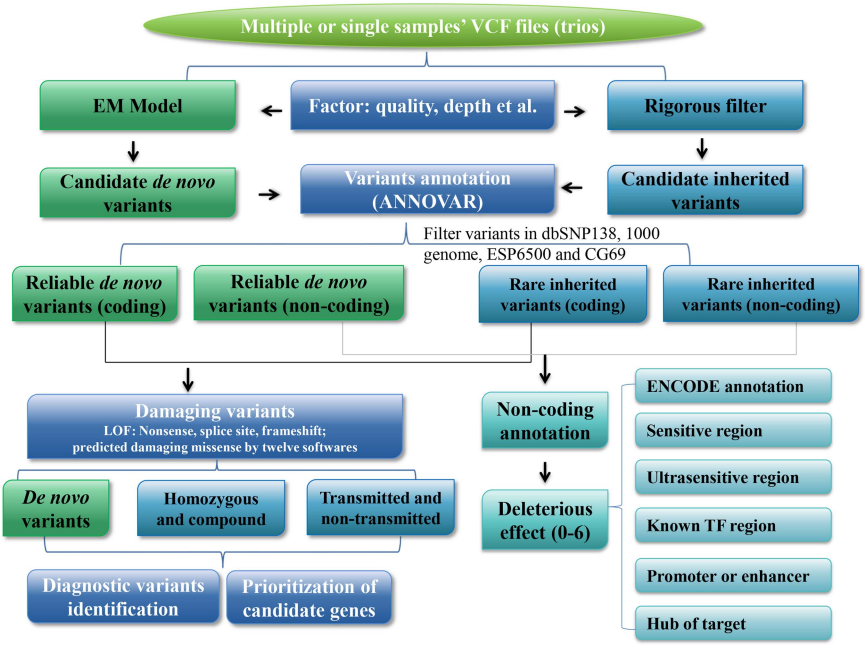


Figure 1 The workflow of mirTrios. mirTrios embarks on the analysis of multiple or single trios-based variant call files (VCF). The workflow and results of mirTrios comprise flowing parts: (1) detection of de novo mutations (DNMs) based on expectation-maximisation (EM) modules; (2) detection of inherited mutations based on rigorous filter; (3) comprehensive annotation of detected variations; (4) detection of diagnostic variants and prioritisation of candidate genes based on annotated extreme mutation; and (5) non-coding annotation and its deleterious effect, considering the areas where they are located.

(variant confidence/quality by depth), MQ0 (number of reads with mapping quality equal to 0), PL (the maximum Phred-scaled likelihoods for genotypes in either parents or child), BT (depth of child/depth of parents), PRT (the maximum percent of the covered reads in proband with reference calls), and PART (the minimum percent of the covered reads in parents with reference calls) are closely relevant to DNMs. Among these properties, PL, PRT and PART are related to family information while the rest are independent from each other. Family information is crucially important to the determination of DNMs, so we also took PL, PRT and PART into inferential account. For those related individuals, we adopted the Bayesian framework to classify them:

$$p\text{ value} \propto P(p_C|p_M,p_F) = \frac{P(p_C,p_M,p_F)}{P(p_M,p_F)}\frac{P(p_C,p_M)P(p_C,p_F)}{\prod_{i\in(C,M,F)}P(p_i)}$$

In the formula,

$$P(p_i) = \begin{cases} PL_i * (1-PRT) & i = C \\ PL_i * (1-PART_i) & i \in (M,F) \end{cases},$$

C refers to proband, F to father and M to mother. $P(p_C,p_M)$, $P(p_C,p_F)$ denote the probability of concurrence of maternal homozygous and proband heterozygous, paternal homozygous and proband heterozygous, respectively. The probability of the proband being heterozygous and the parents being homozygous, which is required for the accurate detection of DNMs, can be obtained by this Bayesian framework.

We used the DNMs validated by Sanger sequencing to build the training set for our EM model. The training set containing both true positive and negative DNMs were extracted from previously published 32 ASD trios[25] and WGS datasets of our in-house, unpublished, 32 other ASD trios. These data were used to generate the initial values in the EM module (such as n, π_i, μ_{z_i} and the variance) for each of the DNM related properties sourced from VCF results. In particular, n refers to the two different Gaussian distributions, DNMs and non-DNMs; μ_{z_i} refers to the mean value of each of the properties (such as QD, MQ0, PL, etc) in the training data. In addition, the initial weight π_i was assigned equally at the first time of the iterative process.

Detection of rare inherited mutations

mirTrios identifies inherited mutations directly from trio-based VCF outputs generated by GATK based on the Phred-scaled probability score and reads depth with related high sensitivity and accuracy.[9 10] The inherited models of mutations were classified into four types according to the genotypes: homozygous or compound heterozygous mutation (Hom), X-linked hemizygous mutation in male (Hem), transmitted heterozygous mutation (THet), and non-transmitted heterozygous mutation (NHet). The four inherited models cause disruption of genes at different levels. Hom affects all copies of genes; Hem disrupts the only copy of genes on the X chromosome in males; while THet and NHet implicate only one copy of genes in the proband and parents, respectively. In addition, mirTrios removed all common mutations by user defined frequency threshold in dbSNP137, ESP6500,[26] and 1000 Genomes (released in April 2012)[27] (figure 1).

Annotation of variants

mirTrios employs ANNOVAR[28] to annotate DNMs and rare inherited mutations with RefSeq (hg19, from UCSC). The annotation information of mutations contains the locations in different genomic regions (exonic, intronic, splicing, intergenic, etc) and the effects on protein coding in coding region (stopgain, frameshift, synonymous, missense, etc). Loss-of-function (LoF) mutations (stopgain, stoploss and splicing single nucleotide variants (SNVs) and frameshift indels) were directly used to prioritise disease candidate genes. Moreover, genes harbouring only synonymous SNVs or non-frameshift indels which were less likely to contribute to disease were eliminated from our candidate list. For non-synonymous SNVs, though many methods or tools have been developed to predict the degree of damages based on evolutionary conservation and functional disruption, all of them have inevitable limitations and biases. A proposed solution for this is to use consensus prediction or majority vote of many methods.[29 30] To this end, mirTrios integrates 12 methods for functional prediction, namely SIFT (Sorting Intolerant from Tolerant),[31] Polyphen2_hvar,[32] Polyphen2_hdiv,[32] MutationTaster,[33] MutationAssessor,[34] LRT,[35] FATHMM (Functional Analysis through Hidden Markov Models),[36] GERP++ (Genomic Evolutionary Rate Profiling),[37] PhyloP,[38] SiPhy,[39 40] RadialSVM and MetaLR in dbNSFP.[29 30] Users can define which of these 12 methods to be used to set pathogenicity thresholds (figure 1).

Prioritisation of candidate genes

Since both de novo and rare inherited mutations are strongly associated with sporadic diseases,[20–23] integrating both of them can be a highly effective way to prioritise candidate genes. TADA incorporates de novo mutations and rare transmitted/non-transmitted heterozygous mutations and adopts parameters for allele frequencies and gene-specific penetrance for risk gene identification.[24] However, LoF/damaging homozygous, compound heterozygous and hemizygous mutations are not taken into account in the primary TADA model. To enrich the prediction model, mirTrios made minor adjustments to the TADA model,[24] and serves to make more accurate predictions of candidate genes, assuming that the effects of those three mutations are equal. The slightly adjusted TADA programme was used to calculate the p value of each gene harbouring rare LoF or damaging mutations (ie, extreme mutations) with statistical support (figure 1).

Non-coding region analysis

Currently, an increasing number of studies have demonstrated that the non-coding regions play important roles in gene regulation, RNA processing, and biological networks.[41] Mutations in the non-coding region have been demonstrated to be associated with many diseases. Therefore, mirTrios supplies de novo and rare inherited mutation annotation in non-coding elements by FunSeq[41] to discover candidate disease drivers with the selected annotation information integrated in this tool. These selected non-coding regions were classified into six functional categories including ENCODE annotation, sensitive region, ultrasensitive region, known transcription factor motif, promoter or enhancer of target genes, and hub of target. The de novo and rare inherited mutations will be assigned a score ranging from 0 to 6, corresponding to its location at different regions, to prioritise non-coding variants (figure 1).

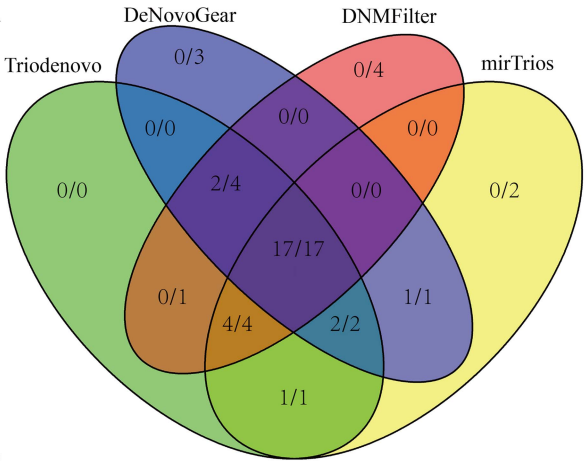
RESULTS

Assessment of identification of DNMs

We tested the performance of mirTrios with WES datasets of our in-house 20 case–parent trios with sporadic ASDs or high myopia generated by WES. We jointly used mirTrios, PolyMutt, DeNovoGear, DNMFilter, and Triodenovo (http://genome.sph.

umich.edu/wiki/Triodenovo) to identify DNMs and totally generate 45 predicted DNMs in coding regions, 27 of which are true positive validated by Sanger sequencing (figure 2A, see online supplementary materials and methods, supplementary table S1). We also compared the accuracy of single-sample calling and multi-sample calling by GATK. Single-sample calling identified 31 more predicted DNMs, but none of them are true positive (figure 2B). By contrast, multi-sample calling has a higher specificity (50% vs 35.5%), yet is still lower than other methods, which suggests that multi-sample calling greatly increases the power of inferring genotypes and haplotypes. Despite a 96.3% sensitivity, the low specificity is a remaining problem for detection of DNMs. Therefore a specialised tool for DNM calling is required. mirTrios detected 27 putative DNMs, 25 of which are true positive, presenting higher sensitivity and specificity (both 92.6%) than PolyMutt, denovoGear and DNMFilter. Triodenovo has a somewhat higher sensitivity (96.3%), but lower specificity (89.7%) than mirTrios. Moreover, mirTrios provides a web-based interface for DNMs and rare inherited mutation detection and candidate gene prioritisation (figure 2B). For the 27 true positive DNMs, all of them were detected by at least two tools and 17 were in the intersection of all four tools. In addition, for the other negative calls, most of them are detected only by one tool. These results indicate that mirTrios achieved a relatively high sensitivity and specificity for detection of DNMs (figure 2B).

A



B

Tools	DNMs (specificity and sensitivity)	Rare inherited	Annotation of mutations	Candidate genes
GATK single-sample	26/76=34.2%; 26/27=96.3%	N	N	N
GATK multi-sample	26/52=50.0%; 26/27=96.3%	Y	N	N
PolyMutt	22/31=70.1%; 22/27=81.5%	Y	N	N
DeNovoGear	22/27=81.5%; 22/27=81.5%	N	N	N
DNMFilter	23/30=76.7%; 23/27=85.2%	N	N	N
Triodenovo	26/29=89.7%; 26/27=96.3%	N	N	N
mirTrios	25/27=92.6%; 25/27=92.6%	Y	Y	Y

Figure 2 Performance comparison of software tools for de novo mutation (DNM) detection. (A) Venn diagram of the detected DNMs from four tools: Triodenovo, DeNovoGear, DNMFilter, and mirTrios. Each part of the Venn diagram represents the counts of true positive DNMs and totally detected DNMs, respectively. (B) Comparison of sensitivity and specificity in the seven tools. mirTrios also supports rare inherited mutation detection, comprehensive annotation, and candidate genes prioritisation.

To provide a guidance for users and define the optimal parameter values for DNM detection by mirTrios, we generated a large amount of simulated data and compared the detection results using a range of parameters (see online supplementary materials and methods). Results showed that some parameters do have an effect on the specificity and sensitivity of DNM detection (see online supplementary figure S1). Based on our simulated data, mirTrios provide an optimal value for each parameter with both high specificity and sensitivity (see online supplementary materials and methods).

Data inputs

In order to facilitate the use of our tools for clinicians lacking sufficient bioinformatics skills, mirTrios provides an intuitive interface to allow user-defined options to customise detection and annotation of de novo and rare inherited mutations generated by trios-based NGS in sporadic disease (figure 2B). Based on these detected mutations and extensive annotation, mirTrios also supports prioritisation of candidate genes. A VCF format file (V.4) generated by GATK and a family list file containing the genetic relationship in each nuclear family are required for mirTrios input (figure 3A). To reduce the input size, all input VCF format files can be compressed into .tar, .gz, .tar.gz, or .tar.bz2 formats. mirTrios allows users to effectively upload the VCF files via the web page or file transfer protocol (FTP) server. After successfully uploading the data, users could start analysis with customised parameters by which the efficiency of the detection of DNMs and rare inherited mutations and annotations could be effectively managed. More importantly, this flexible customisation enables users to re-analyse uploaded data independently through different combinations of parameters.

To make mirTrios more convenient, the mirTrios stand-alone version supports BAM files as inputs. Public users can download this freely available stand-alone program from the mirTrios website. Since the size of a BAM file is generally 100-fold greater than that of a VCF file (eg, the size of BAM and VCF files of an exome are 5 GB and 50 MB, respectively), which is a stumbling block for uploading, the web server of mirTrios will only support VCF files as input. However, it is noted that mirTrios is specifically developed for comprehensive analysis of sporadic diseases instead of familiar diseases, such as three generation families. It is considered that familiar diseases are generally used to identify rare inherited variations, which are supposed to segregate with disease, rather than DNMs. Therefore, the current version of mirTrios only works on nuclear families with multiple probands and/or siblings and their unaffected parents.

Data outputs

The analytical results can be retrieved and browsed by a unique identifier which is generated immediately after the data are uploaded successfully (figure 3A). A typical output includes four sections: DNMs and annotation; rare inherited mutations and annotation; disease candidate genes; and non-coding region analysis (figure 3B–E). These four sections are well organised to demonstrate the results of each part. The first section illustrates all the detected DNMs and annotations of them, including mutation loci (exonic, splicing, 5′UTR, upstream, etc), mutational type (SNV, insertion and deletion), and effects on coding region (stoploss, stopgain, non-synonymous, synonymous, frameshift, etc), as well as the annotation in various public databases, such as dbSNP138, ESP6500,[26] and 1000 Genomes.[27] For non-synonymous SNVs, mirTrios provides a predicted pathogenicity score based on 12 methods, which can be modified electively. More importantly, mirTrios also supports the

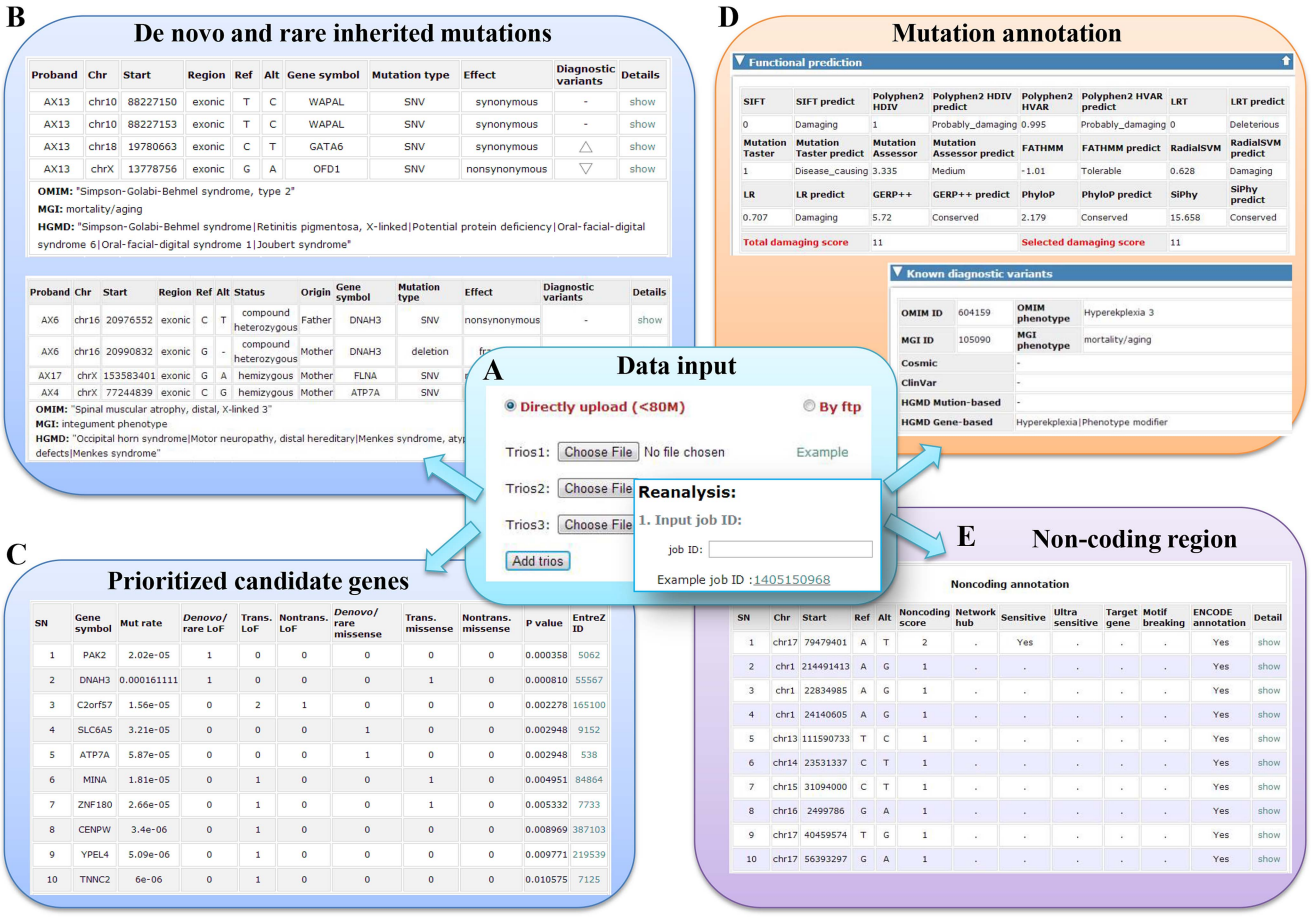


Figure 3 The snapshot of the results of mirTrios. (A) Trios-based variant call files (VCF) and family list are loaded as input along with several user-selected options. (B) Detected de novo mutations (DNMs) and rare inherited mutations. (C) Comprehensive annotation of detected variations. (D) Known diagnostic variant identification and candidate gene prioritisation. (E) Estimates of the deleterious effect of variations in the non-coding region.

detection of known diagnostic mutations and disease-related genes based on five resources: OMIM (Online Mendelian Inheritance in Man),[42] MGI,[43] HGMD (Human Gene Mutation Database),[44] COSMIC (Catalogue of Somatic Mutations in Cancer),[45] and ClinVar.[46] This is powerful for the identification of known functional mutations and novel candidate genes. The second section showed all classes of detected rare inherited mutations (homozygous or compound heterozygous, X-linked hemizygous, and transmitted/non-transmitted heterozygous mutations) with detailed annotation similar to DNMs. The disease candidate genes section displays all the potential disease-associated genes, which contain at least one extreme mutation (damaging de novo or rare inherited mutation) with a given p value. In this section, mirTrios clearly provide the count of LoF/damaging DNMs, and transmitted/non-transmitted rare inherited mutations in each gene. The optional non-coding region analysis results will be generated if users provide the whole genome trios sequencing data. In this section, mirTrios shows all detected de novo and rare inherited mutations located in the functional non-coding region. Based on the sequence location, mirTrios provides a score ranging from 0 (less deleterious) to 6 (more deleterious) to estimate the deleterious effect of variations.

DISCUSSION

The rapid advances of WES/WGS technologies have greatly facilitated clinical genetic diagnosis genome-wide.[47 48] For sporadic disease, despite the minor role of common mutations or the environment, LoF/damaging DNMs is an important source of causality.[49] In addition, rare inherited mutation also contributes to the risk of sporadic disease, such as ASD[22] and schizophrenia.[50] The vast amount of mutations generated by NGS poses multiple challenges for the identification of functional mutations and candidate genes.

At present, although a few tools have been developed to detect DNMs or candidate genes by NGS, there are still no public online services available for comprehensive analysis of trios-based NGS data. Therefore, we have developed a novel and comprehensive platform, mirTrios, for the analysis of trio-based WES/WGS VCF results, which allows accurate detection and annotation of DNMs and rare inherited mutation in coding and non-coding regions. For the average geneticist and clinician, the integrated framework of mirTrios avoids the cumbersome process of complex installation, redundant operations, and requirement for high-performance computational capability. For multiple trios analysis, mirTrios also provides an integrated framework for known diagnostic variant identification and candidate gene prioritisation based on the detected de novo and rare inherited mutations from the large amount of data generated by NGS. In essence, mirTrios provides comprehensive and meaningful data for users to study in depth the genetic basis of sporadic diseases.

mirTrios provides an intuitive interface for users to upload files directly by web page or ftp address, which can be widely used by researchers to explore the functional mutation and

candidate genes in sporadic disease. mirTrios is freely available for non-commercial use and will be updated regularly to keep up with the latest resources of the implemented databases. Restricted by the lack of a sophisticated algorithm for detecting de novo CNV and SV (structural variation), mirTrios currently only provides point mutation analysis. In this aspect, mirTrio will be updated with state-of-art de novo CNV/SV detection and integrate these tools with optimal accuracy and specificity. We believe mirTrios will be very helpful for the study of sporadic disease.

IMPLEMENTATION

mirTrios is freely available at http://centre.bioinformatics.zj.cn/mirTrios/. Documentation and example data can be found on the website. The web client of mirTrios was implemented independently and has been successfully tested with different releases of Microsoft Internet Explorer 11.0, Firefox 30.0, Google Chrome 35.0, and Safari 5.1 (under different versions of MacOS, Microsoft Windows and Linux). mirTrios was constructed under an Apache/PHP/MySQL environment on the Red Hat Enterprise 5.5 Linux operating system. The uploaded VCF data will be analysed on our five computational nodes, with 16 CPUs and 32 GB of RAM in each node.

Acknowledgements The authors thank Dr Yong-hui Jiang and Dr Ming-bang Wang for helpful training data support.

Contributors JW, ZSS and KX: designed and supervised the study. JL, YJ, TW, and HC drafted the manuscript. JL, YJ, TW, QS, and XR: developed the web server. YJ, TW, QX, and JL: developed the method to detect DNMs. JL: implemented the method to detect rare inherited mutations and prioritisation of candidate genes. All the authors read and approved the manuscript.

Funding The project was funded by the National Basic Research Program of China (No. 2012CB517902 and 2012CB517904), the National "12th Five-Year" scientific and technological support projects (No. 2012BAI03B02), and the Special Funds of National Health and Family Planning Commission of China (No. 201302002).

Competing interests None.

Patient consent Obtained.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Datasets used in the manuscript are available on the mirTrios web server.

REFERENCES

1 Ku C, Polychronakos C, Tan E, Naidoo N, Pawitan Y, Roukos D, Mort M, Cooper D. A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. *Mol Psychiatry* 2012;18:141–53.
2 Gratten J, Visscher PM, Mowry BJ, Wray NR. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat Genet* 2013;45:234–8.
3 Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nat Rev Genet* 2012;13:565–75.
4 Hoischen A, Krumm N, Eichler EE. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat Neurosci* 2014;17:764–72.
5 Stessman HA, Bernier R, Eichler EE. A genotype-first approach to defining the subtypes of a complex disease. *Cell* 2014;156:872–7.
6 Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, Romano-Adesman A, Bjornson RD, Breitbart RE, Brown KK. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 2013;498:220–3.
7 Krumm N, O'Roak BJ, Shendure J, Eichler EE. A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci* 2014;37:95–105.
8 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–9.
9 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
10 Chen W, Li B, Zeng Z, Sanna S, Sidore C, Busonero F, Kang HM, Li Y, Abecasis GR. Genotype calling and haplotyping in parent-offspring trios. *Genome Res* 2013;23:142–51.
11 Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009;25:2283–5.
12 Peng G, Fan Y, Palculict TB, Shen PD, Ruteshouser EC, Chi AK, Davis RW, Huff V, Scharfe C, Wang WY. Rare variant detection using family-based sequencing analysis. *Proc Natl Acad Sci USA* 2013;110:3985–90.
13 Santoni FA, Makrythanasis P, Nikolaev S, Guipponi M, Robyr D, Bottani A, Antonarakis SE. Simultaneous identification and prioritization of variants in familial, de novo, and somatic genetic disorders with VariantMaster. *Genome Res* 2014;24:349–55.
14 Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014;46:818–25.
15 Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, Cucca F, Kang HM, Abecasis GR. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet* 2012;8:e1002944.
16 Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, Conrad DF. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods* 2013;10:985–7.
17 Liu Y, Li B, Tan R, Zhu X, Wang Y. A gradient-boosting approach for filtering de novo mutations in parent-offspring trios. *Bioinformatics* 2014;30:1830–6.
18 Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 2012;151:1431–42.
19 Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee Y-h, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat Methods* 2014;11:1033–6.
20 Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, Neale BM, Kirby A, Ruderfer DM, Fromer M. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* 2013;77:235–42.
21 Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, Schmitz-Abe K, Harmin DA, Adli M, Malik AN. Using whole-exome sequencing to identify inherited causes of autism. *Neuron* 2013;77:259–73.
22 Stein JL, Parikshak NN, Geschwind DH. Rare inherited variation in autism: beginning to see the forest and a few trees. *Neuron* 2013;77:209–11.
23 Toma C, Torrico B, Hervás A, Valdés-Mas R, Tristán-Noguero A, Padillo V, Maristany M, Salgado M, Arenas C, Puente X. Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Mol Psychiatry* 2014;19:784–90.
24 He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* 2013;9:e1003671.
25 Jiang Y-h, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet* 2013;93:249–63.
26 Fu WQ, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, Akey JM; NHLBI Exome Sequencing Project. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants (vol 493, pg 216, 2013). *Nature* 2013;495:270.
27 Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491:56–65.
28 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
29 Liu X, Jian X, Boerwinkle E. dbNSFP v2. 0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* 2013;34:E2393–402.
30 Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* 2011;32:894–9.
31 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–81.
32 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–9.
33 Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;7:575–6.
34 Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;39:e118.
35 Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009;19:1553–61.
36 Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013;34:57–65.
37 Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;6:e1001025.

38 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel KA. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome res* 2010;20:110–21.
39 Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 2009;25:i54–62.
40 Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 2011;478:476–82.
41 Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 2013;342:1235587.
42 Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;33(Database issue):D514–517.
43 Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE. The Mouse Genome Database Group. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* 2014;42:D810–D81.
44 Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinformatics* 2012;Chapter 1: Unit1.13. http://dx.doi.org/10.1002/0471250953.bi0113s39
45 Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011;39(Database issue): D945–50.
46 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;42(Database issue):D980–5.
47 MacArthur D, Manolio T, Dimmock D, Rehm H, Shendure J, Abecasis G, Adams D, Altman R, Antonarakis S, Ashley E. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;508:469–76.
48 Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *N Engl J Med* 2014;370:2418–25.
49 Ronemus M, Iossifov I, Levy D, Wigler M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet* 2014;15:133–41.
50 Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'Dushlaine C, Chambert K, Bergen SE, Kähler A. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 2014;506:185–90.

mirTrios: an integrated pipeline for detection of de novo and rare inherited mutations from trios-based next-generation sequencing

Jinchen Li, Yi Jiang, Tao Wang, Huiqian Chen, Qing Xie, Qianzhi Shao, Xia Ran, Kun Xia, Zhong Sheng Sun and Jinyu Wu

J Med Genet published online January 16, 2015

Gene 519 (2013) 18–25

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Genome-wide identification and divergent transcriptional expression of StAR-related lipid transfer (START) genes in teleosts

Huajing Teng [a,c,1], Wanshi Cai [a,b,1], Kun Zeng [c], Fengbiao Mao [a,b], Mingcong You [c], Tao Wang [c], Fangqing Zhao [a,*], Zhongsheng Sun [a,c,**]

[a] Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China
[b] Graduate University of the Chinese Academy of Sciences, Beijing 100049, China
[c] Institute of Genomic Medicine, Wenzhou Medical College, Wenzhou 325035, China

ARTICLE INFO

Article history:
Accepted 30 January 2013
Available online 13 February 2013

Keywords:
START genes
Gene loss
Fish-specific genome duplication
Transcriptional expression pattern

ABSTRACT

The lipid transfer reactions and the steroidogenic acute regulatory protein (StAR)-related lipid transfer (START) genes have a major role in lipid metabolism. However, START genes and their physiological functions in teleost fishes are relatively unknown. Through genome-wide screening, we identified and annotated 91 START genes in 5 teleost species. Although START domain-containing proteins are augmented in teleost genomes relative to tetrapod genomes, a similar number of genes are shared between them. Asymmetry of paralogous gene loss within the teleost START family and an extra copy of some START genes in teleosts resulting from fish-specific genome duplication have been demonstrated. A distinct transcriptional expression pattern within members of some START groups under different developmental stages suggests divergent functions within the same group in the developmental process. In addition, an asymmetric molecular evolution rate deviating from the neutral expectation has been observed in 7 of 14 teleost fish extra-duplicated pairs. The present study provides valuable information for increasing our understanding of the evolution and gene expression divergence under developmental stages of the START gene family in teleost fishes.

1. Introduction

Gene duplication, which has long been regarded as one of the major forces of evolution, can facilitate the acquisition of new functions of duplicated genes through neo-functionalization (Ohno, 1970) or partitioning of the ancestral gene functions between descendant duplicated genes by sub-functionalization (Conant and Wolfe, 2008; Force et al., 1999; He and Zhang, 2005; Lynch and Force, 2000). Tandem duplication, segmental duplication and whole-genome duplication (WGD) are major gene duplication mechanisms in eukaryotes. WGD is particularly intriguing because it has been regarded as a parsimonious evolutionary innovation of gene duplication (Haldane, 1932; Ohno, 1970; Taylor and Raes, 2004). It is well accepted that fish-specific genome duplication (FSGD) occurred prior to the teleost radiation (Amores et al., 1998; Christoffels et al., 2004; Jaillon et al., 2004; Teng et al., 2010; Vandepoele et al., 2004; Woods et al., 2005). However, the total number of genes in teleost species is not twice that present in tetrapod species, which prompted us to investigate the gene loss event after FSGD and to investigate whether fish-specific duplicated paralogs evolve at similar rates after duplication. As the results of various gene duplication and loss events, gene families provide a unique source for studying the evolutionary relationships between genes both within and between organisms. Changes in family size due to lineage-specific gene duplication or loss might provide insights into the evolutionary forces that have shaped eukaryotic genomes (Demuth et al., 2006). Thus, inferring an evolutionary scenario for a gene family is essential to understanding the phenotypic diversification of eukaryotic organisms (Hanada et al., 2009; Sato et al., 2009).

The steroidogenic acute regulatory protein (StAR)-related lipid transfer (START) domain, named after the mammalian 30 kDa StAR protein, is a protein module of around 200 amino acids implicated in the control of several aspects of lipid biology, including lipid trafficking, lipid metabolism and cell signaling (Alpy and Tomasetto, 2005; Soccio and Breslow, 2003). Mutation or misexpression of some START proteins was also reported to link to some pathological processes, including genetic disorders, autoimmune disease and cancer (Alpy and Tomasetto, 2005). Members of the START domain family have been shown to bind different ligands, such as sterols (e.g., StAR or STARD1) and lipids (e.g., PCTP or

Abbreviations: START, steroidogenic acute regulatory protein-related lipid transfer; WGD, whole-genome duplication; FSGD, fish-specific genome duplication; PH, pleckstrin homology; SAM, sterile alpha motif; RhoGAP, Rho-type GTPase-activating protein; 4HBT, 4-hydroxybenzoate thioesterase; HMM, hidden Markov model; d_N, nonsynonymous substitutions per nonsynonymous site; d_S, synonymous substitutions per synonymous site; RTK, receptor tyrosine kinase; Dr, *Danio rerio*; Tr, *Takifugu rubripes*; Tn, *Tetraodon nigroviridis*; Ol, *Oryzias latipes*; Ga, *Gasterosteus aculeatus*; Lc, *Latimeria chalumnae*; Xt, *Xenopus tropicalis*; Ac, *Anolis carolinensis*; Gg, *Gallus gallus*; Mm, *Mus musculus*; Hs, *Homo sapiens*; Ci, *Ciona intestinalis*.

* Corresponding author. Tel.: +86 10 64869325; fax: +86 10 64880586.
** Correspondence to: Z. Sun, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China. Tel.: +86 10 64864959; fax: +86 10 64880586.
E-mail addresses: biofqzhao@gmail.com (F. Zhao), zsunusa@yahoo.com (Z. Sun).
[1] These authors contributed equally to this work.

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.01.058

STARD2), and exhibit enzymatic activity. Some other functional domains that were found associated with START in animals include pleckstrin homology (PH), sterile alpha motif (SAM), Rho-type GTPase-activating protein (RhoGAP), and 4-hydroxybenzoate thioesterase (4HBT) (Schrick et al., 2004; Soccio and Breslow, 2003). Ligand binding by START domain can regulate the activity of other domains within multi-domain proteins, such as the RhoGAP domain, the homeodomain and the thioesterase domain (Iyer et al., 2001; Ponting and Aravind, 1999). START domain is evolutionarily conserved in plants and animals. Fifteen START domain-containing proteins (STARD1–STARD15) have been identified in humans (Soccio and Breslow, 2003), and hundreds have been determined in invertebrates, bacteria and plants. However, only a very few START homologs have been reported in teleost fishes, which are the largest and most diverse group of vertebrates. The availability of sequenced and assembled genomes of zebrafish (*Danio rerio*), fugu (*Takifugu rubripes*) (Aparicio et al., 2002), Green Spotted Puffer (*Tetraodon nigroviridis*) (Jaillon et al., 2004), medaka (*Oryzias latipes*) (Kasahara et al., 2007) and three-spined stickleback (*Gasterosteus aculeatus*) has provided an opportunity for the genome-wide screening of START homologs and comparative analysis in teleost fish species.

In this study, we identified and annotated the START gene family members in teleosts through genome-wide screening and we investigated their transcript expression profile under experimental conditions, domain composition and phylogenetic relationships. In addition, a relative rate test was used to examine whether one of the duplicates has evolved at an accelerated rate following the duplication. With such an in-depth investigation, we expected to provide a detailed case in studying how genes evolve after gene duplication, and provide some data for future physiological function research of START genes in fishes.

2. Materials and methods

2.1. Data sets and phylogenetic analysis

Fifteen human START proteins were used as the query sequences in BLASTP and TBLASTN searches ($E<1e^{-5}$) against the NCBI (Maglott et al., 2010) or Ensembl databases (March 2011) (Flicek et al., 2011) of human, mouse, chicken, anole lizard, Western clawed frog, coelacanth, zebrafish, medaka, Green Spotted Puffer, three-spined stickleback, fugu and Sea squirts. Each matching sequence was used iteratively to search the databases until no new sequence was found. Additionally, a hidden Markov model (HMM) search (Johnson et al., 2010) was done in the proteome databases of the species listed above using the START domain (PFAM, PF01852). All protein sequences derived from the collected candidate START genes were further examined using the PFAM program (Mistry and Finn, 2007) with the default cut-off parameters. The amino acid sequence alignment of START domains was generated using MUSCLE (Edgar, 2004) with the default setting. A bootstrap consensus phylogenetic tree was constructed using the maximum-likelihood method in MEGA5 (Tamura et al., 2011) under the JTT + G model.

2.2. Analysis of synteny

All predicted genes within 20 Mb of each human or mouse START paralog were obtained using the BioMart mode in Ensembl (March 2011). Genes exhibiting orthologous relationship in both species (human/mouse) and supported by phylogenetic analysis were selected for syntenic analysis. Neighboring genes flanking the chicken, anole lizard, Western clawed frog, zebrafish, medaka, fugu, Green Spotted Puffer or three-spined stickleback START paralogs were obtained using the BioMart mode in Ensembl from dataset of the chicken (WASHUC2), anole lizard (AnoCar2.0), Western clawed frog (JGI_4.2), zebrafish (Zv9), medaka (MEDAKA1) fugu (FUGU4), Green Spotted Puffer (TETRAODON8) or three-spined stickleback (BROADS1) genome, respectively. Blocks of synteny were constructed on the basis of the orthologous relationship of genes among different species.

2.3. Transcriptional expression analysis of teleost START genes

The genome-wide microarray data of zebrafish (Domazet-Loso and Tautz, 2010), medaka (Iwahashi et al., 2009), Green Spotted Puffer (Chan et al., 2009) and three-spined stickleback were obtained from the NCBI Gene Expression Omnibus (GEO) with accession numbers GSE24616, GSE15380, GSE12976 and GSE34783, respectively. Extraction and filtration of each microarray data were processed as previously described (Chan et al., 2009; Domazet-Loso and Tautz, 2010; Iwahashi et al., 2009). Probe sets corresponding to the putative zebrafish STARTs were identified from Agilent Zebrafish (V2) Gene Expression Microarrays, NimbleGen Oryzias latipes_TIGR_rel5 (GFC023) 27 k array, Agilent custom 44 K Tetraodon array or Agilent-016492 three-spined stickleback 44 K 60 nt oligo array version 1.0. If more biological replicates were used in a specific experiment, such as 2–4 replicates in zebrafish, 5 replicates in three-spined stickleback and 3 replicates in medaka, mean of the expression values among the replicates were used. For genes with more than one set of probes, the mean of expression values was considered. Finally, the $\log_2$ transformed transcript intensity data were hierarchically clustered on the basis of the Euclidean distance with complete linkage in the Cluster program (de Hoon et al., 2004), and the relative transcript accumulation was represented in a color code with green or red showing the lower or higher levels of transcriptional expression, respectively.

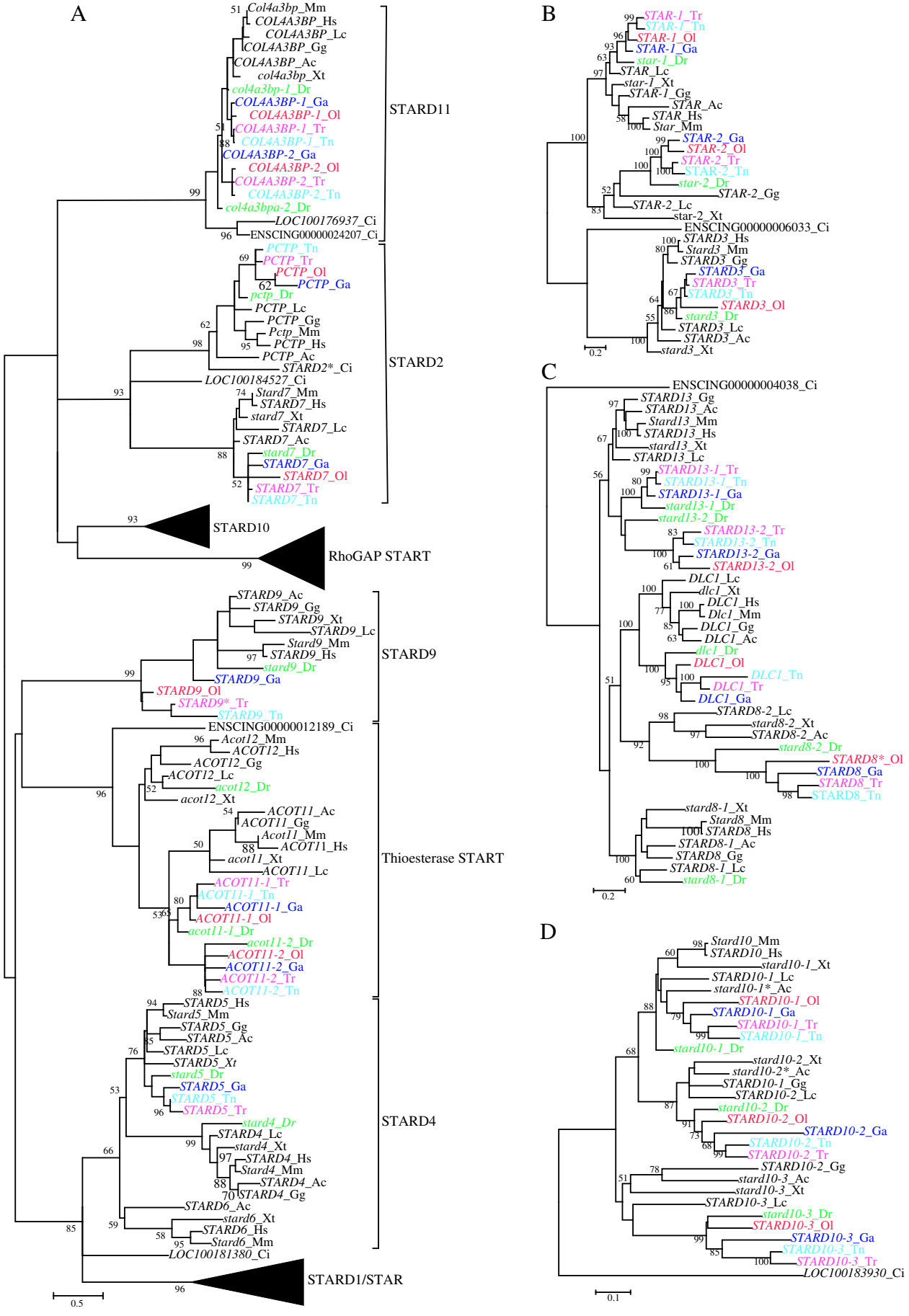
2.4. Divergence and relative rate test of duplicated teleost START pairs

The coding sequences of the duplicated teleost START pairs were aligned following the amino acid alignment by CodonAlign 2.0 (http://homepage.mac.com/barryghall/CodonAlign.html). Pairwise calculation of d_N/d_S between these teleost START pairs is estimated with the yn00 program of PAML4 (Yang, 2007). Further, nonparametric relative rate tests were done with amino acid sequences to investigate whether one of these teleost START pairs has evolved at an accelerated rate following the duplication using MEGA (Tamura et al., 2011). To test whether some sites were under positive selection, several site-specific models (M0, M1, M2, M3, M7 and M8) and branch site test 2 were used to detect positive selection using the codeml program implemented in PAML4 (Yang, 2007).

3. Results

3.1. Identification and phylogenetic analysis of START genes

Through extensive similarity-based searches, we identified 91 teleost START genes: 18 in Green Spotted Puffer, three-spined stickleback or fugu, 21 in zebrafish, and 16 in medaka (Supplemental Table 1). The length of STARTs in teleosts ranged from 198 to 1928 amino acid residues, and the number of exons ranged from 5 to 19 (Supplemental Table 1 and Fig. 1). In teleosts, about half of the START domain-containing proteins (48/91) are multi-domain proteins. Functional domains associated with START in teleosts include pleckstrin homology (PH) in STARD11s(COL4A3BPs), sterile alpha motif (SAM), Rho-type GTPase-activating protein (RhoGAP) in STARD8s, STARD12s(DLC1s) and STARD13s, and 4-hydroxybenzoate thioesterase (4HBT) in STARD14s(ACOT11s), consistent with earlier reports for non-teleosts (Alpy and Tomasetto, 2005; Schrick et al., 2004; Soccio and Breslow, 2003) (Supplemental Fig. 1). Although START domain-containing proteins are augmented in teleost genomes relative to mammalian genomes (Supplemental Table 1), similar gene numbers are found for each. In order to investigate the evolutionary relationship among these teleost START genes, 91 teleost START genes, 18 coelacanth and 80 tetrapod START orthologs, and 9 ascidian START genes were used in phylogenetic analysis by the maximum-likelihood method in MEGA5. The analysis unambiguously defined 8 START groups with high bootstrap values (Fig. 1); namely, the STARD1/STAR, STARD4, RhoGAP START, STARD2, STARD10, STARD11, thioesterase START and STARD9 group. The gene number of


A
Col4a3bp_Mm
COL4A3BP_Hs
COL4A3BP_Lc
COL4A3BP_Gg
COL4A3BP_Ac
col4a3bp_Xt
col4a3bp-1_Dr
COL4A3BP-1_Ga
COL4A3BP-1_Ol
COL4A3BP-1_Tr
COL4A3BP-1_Tn
COL4A3BP-2_Ga
COL4A3BP-2_Ol
COL4A3BP-2_Tr
COL4A3BP-2_Tn
col4a3bpa-2_Dr
LOC100176937_Ci
ENSCING00000024207_Ci
STARD11
PCTP_Tn
PCTP_Tr
PCTP_Ol
PCTP_Ga
pctp_Dr
PCTP_Lc
PCTP_Gg
Pctp_Mm
PCTP_Hs
PCTP_Ac
STARD2*_Ci
LOC100184527_Ci
Stard7_Mm
STARD7_Hs
stard7_Xt
STARD7_Lc
STARD7_Ac
stard7_Dr
STARD7_Ga
STARD7_Ol
STARD7_Tr
STARD7_Tn
STARD2
STARD10
RhoGAP START
STARD9_Ac
STARD9_Gg
STARD9_Xt
STARD9_Lc
Stard9_Mm
STARD9_Hs
stard9_Dr
STARD9_Ga
STARD9_Ol
STARD9*_Tr
STARD9_Tn
STARD9
ENSCING00000012189_Ci
Acot12_Mm
ACOT12_Hs
ACOT12_Gg
ACOT12_Lc
acot12_Dr
acot12_Xt
ACOT11_Ac
ACOT11_Gg
Acot11_Mm
ACOT11_Hs
acot11_Xt
ACOT11_Lc
ACOT11-1_Tr
ACOT11-1_Tn
ACOT11-1_Ga
ACOT11-1_Ol
acot11-1_Dr
acot11-2_Dr
ACOT11-2_Ol
ACOT11-2_Ga
ACOT11-2_Tr
ACOT11-2_Tn
Thioesterase START
STARD5_Hs
Stard5_Mm
STARD5_Gg
STARD5_Ac
STARD5_Lc
STARD5_Xt
stard5_Dr
STARD5_Ga
STARD5_Tn
STARD5_Tr
stard4_Dr
STARD4_Lc
stard4_Xt
STARD4_Hs
Stard4_Mm
STARD4_Ac
STARD4_Gg
STARD6_Ac
stard6_Xt
STARD6_Hs
Stard6_Mm
LOC100181380_Ci
STARD4
STARD1/STAR
0.5
B
STAR-1_Tr
STAR-1_Tn
STAR-1_Ol
STAR-1_Ga
star-1_Dr
STAR_Lc
star-1_Xt
STAR-1_Gg
STAR_Ac
STAR_Hs
Star_Mm
STAR-2_Ga
STAR-2_Ol
STAR-2_Tr
STAR-2_Tn
star-2_Dr
STAR-2_Gg
STAR-2_Lc
star-2_Xt
ENSCING00000006033_Ci
STARD3_Hs
Stard3_Mm
STARD3_Gg
STARD3_Ga
STARD3_Tr
STARD3_Tn
STARD3_Ol
stard3_Dr
STARD3_Lc
STARD3_Ac
stard3_Xt
0.2
C
ENSCING00000004038_Ci
STARD13_Gg
STARD13_Ac
Stard13_Mm
STARD13_Hs
stard13_Xt
STARD13_Lc
STARD13-1_Tr
STARD13-1_Tn
STARD13-1_Ga
stard13-1_Dr
stard13-2_Dr
STARD13-2_Tr
STARD13-2_Tn
STARD13-2_Ga
STARD13-2_Ol
DLC1_Lc
dlc1_Xt
DLC1_Hs
Dlc1_Mm
DLC1_Gg
DLC1_Ac
dlc1_Dr
DLC1_Ol
DLC1_Tn
DLC1_Tr
DLC1_Ga
STARD8-2_Lc
stard8-2_Xt
STARD8-2_Ac
stard8-2_Dr
STARD8*_Ol
STARD8_Ga
STARD8_Tr
STARD8_Tn
stard8-1_Xt
Stard8_Mm
STARD8_Hs
STARD8-1_Ac
STARD8_Gg
STARD8-1_Lc
stard8-1_Dr
0.2
D
Stard10_Mm
STARD10_Hs
stard10-1_Xt
STARD10-1_Lc
stard10-1*_Ac
STARD10-1_Ol
STARD10-1_Ga
STARD10-1_Tr
STARD10-1_Tn
stard10-1_Dr
stard10-2_Xt
stard10-2*_Ac
STARD10-1_Gg
STARD10-2_Lc
stard10-2_Dr
STARD10-2_Ol
STARD10-2_Ga
STARD10-2_Tn
STARD10-2_Tr
STARD10-2_Gg
stard10-3_Ac
stard10-3_Xt
STARD10-3_Lc
stard10-3_Dr
STARD10-3_Ol
STARD10-3_Ga
STARD10-3_Tn
STARD10-3_Tr
LOC100183930_Ci
0.1

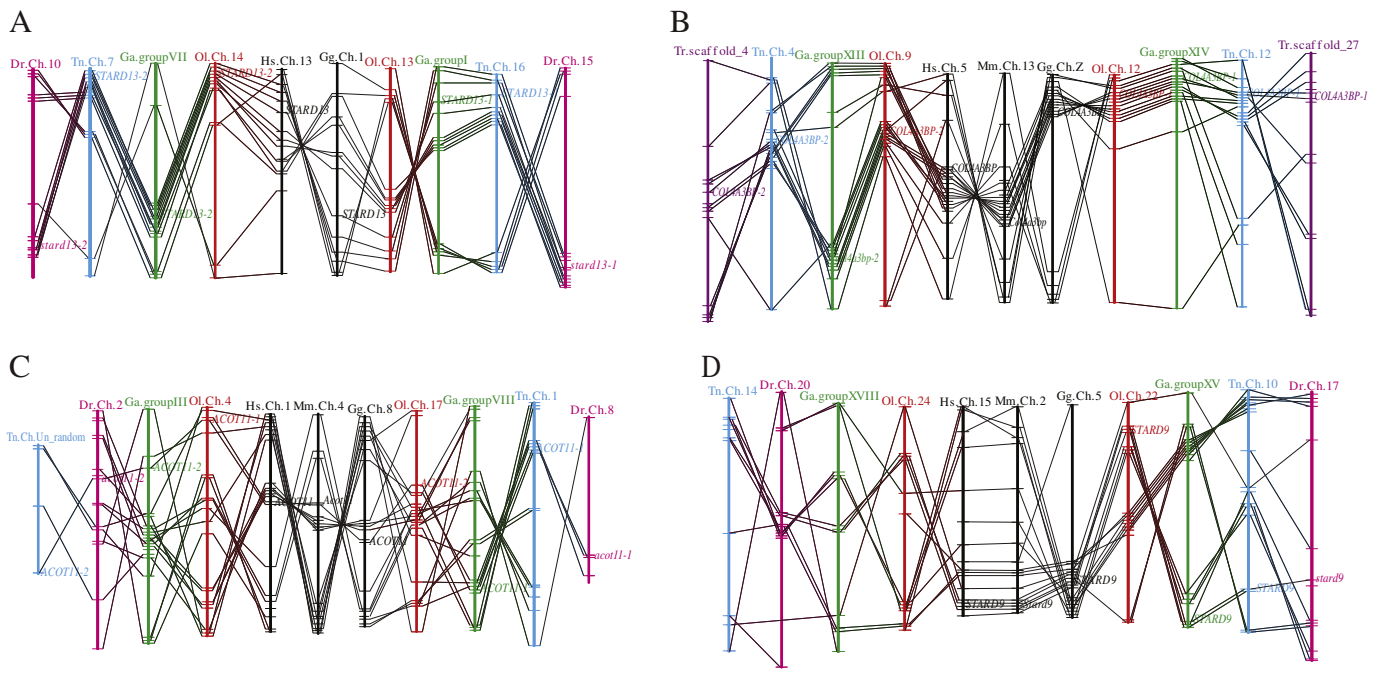


Fig. 2. Extra paralogons surrounding (A) *STARD13* (B) *COL4A3BP*, (C) *ACOT11* and (D) *STARD9* genes in teleost genomes. (A) The tetrapod START13 paralogon has two fish paralogons, the START13-1 paralogon and the START13-2 paralogon. Each of the *STOML3*, *DCLK1*, *SLC7A1*, *TRPC4*, *FREM2*, *SPG20*, *NBEA*, *B3GALTL*, *FRY* and *HMGB1* on human chromosome 13 near *START13* has two co-orthologs in medaka chromosomes 13 and 14, three-spined stickleback groups VII and I, Green Spotted Puffer chromosomes 16 and 7, zebrafish chromosomes 15 and 10. (B) The COL4A3BP-1 and COL4A3BP-2 fish paralogons are accompanied by one tetrapod START11 paralogon. *SV2C*, which is near *COL4A3BP* on human chromosome 5, has two co-orthologs located in medaka chromosomes 12 and 9, three-spined stickleback groups XIII and XIV, Green Spotted Puffer chromosomes 12 and 4, zebrafish chromosomes 5 and 21, fugu scaffolds 4 and 27. (C) Each ACOT11 paralogon defines a synteny with a high degree of conservation among the studied tetrapod and fish species. (D) Although no extra copy of START9 genes was found in some fish genomes, two conserved fish paralogons flanking each gene were observed, suggesting extra duplication of this STARD9 gene that occurred in the teleost ancestor, but was lost before the divergence of zebrafish. The positions of genes on chromosomes are not drawn to scale.

each group varies dramatically, ranging from 11 to 42. Some orthologs of tetrapod START genes, such as *STARD6*, are absent from all teleost fishes and some are absent from a certain fish lineage, such as *STARD4*_Tn and *STARD5*_Ol. Several START genes (*STARD1*, *STARD8*, *STARD10*, *COL4A3BD*, *STARD13* and *ACOT11*) have multiple copies in certain fish genomes, unlike mammalian genomes, indicating that extra rounds of duplication have occurred in the teleost lineage (Fig. 1).

3.2. Syntenic analysis

Orthologous genes flanking each START gene define a syntenic conservation among tetrapods and teleosts (Fig. 2; Supplemental Table 2, Supplemental Fig. 2 and Supplemental Fig. 3). Although some orthologs of tetrapod START genes, such as *STARD6*, are absent from all teleost fish genomes, syntenic regions of *STARD6* genes can be observed in teleosts (Supplemental Fig. 3). The same applies to some fish lineage-specific lost genes, e.g. *STARD4*_Tn, *ACOT12*_Tn, and *STARD*_Ol (Supplemental Fig. 3), suggesting that these genes were present in ancestral teleosts but were lost after the divergence of teleost fishes. Extra copies of some START genes, such as *STARD1*, *STARD8*, *STARD10*, *COL4A3BP*, *STARD13* and *ACOT11*, were observed in some fish genomes compared to mammalian genomes, indicating that extra duplication in the teleost lineage or mammalian-specific gene losses might have occurred in the evolutionary history of these genes (Postlethwait, 2007). Therefore, more tetrapod lineages (chicken, lizard and frog) were used in our synteny analysis to investigate their evolutionary history. Although frequent gene-linkage disruptions, micro-inversions or rearrangements occurred in teleosts, we could find some traces of extra paralogons around each of the STARD1, -8 or -10 genes in teleost and non-mammalian tetrapod genomes (Supplemental Fig. 2). For example, 2 chicken or fish paralogons STAR-1 and STAR-2 are accompanied by 1 mammalian STAR paralogon. It appears that extra paralogs of these 3 genes were generated in the earlier vertebrates before the divergence of the Teleostomi followed by mammalian-specific gene losses. Thus, the difference in gene number of STARD1, -8 or -10 genes between teleosts and mammals is most likely because of mammalian-specific gene losses (Postlethwait, 2007). However, extra paralogons around each of the *COL4A3BP*, *STARD13* and *ACOT11* genes were found in teleost genomes but not in tetrapod genomes. For example, the tetrapod START13 paralogon has 2 fish paralogons, START13-1 and START13-2 (Fig. 2). Each of *STOML3*, *DCLK1*, *SLC7A1*, *TRPC4*, *FREM2*, *SPG20*, *NBEA*, *B3GALTL*, *FRY* and *HMGB1* on human chromosome 13 near START13 has two co-orthologs on medaka chromosomes 13 and 14, three-spined stickleback groups VII and I, Green Spotted Puffer chromosomes 16 and 7 and zebrafish chromosomes 15 and 10, which suggests that extra copies of the COL4A3BP, STARD13 and ACOT11 genes in teleosts might be the results of FSGD (Kasahara et al., 2007). Although no extra copy of the STARD9 gene was found in some fish genomes, 2 conserved fish paralogons flanking each gene were also

Fig. 1. (A) Phylogenetic analysis of the START gene family and detailed phylogeny of (B) STARD1/STAR, (C) RhoGAP START and (D) STARD10. The bootstrap consensus phylogenetic tree was constructed using the maximum-likelihood method in MEGA5, and the numbers indicate the percentage bootstrap support. The symbol '◂' represents the compressed START group of STARD1/STAR, RhoGAP START or STARD10, which are expanded in (B), (C) or (D), respectively. *Gene predicted using FGENESH + software (http://linux1.softberry.com/berry.phtml). Dr, *Danio rerio*; Tr, *Takifugu rubripes*; Tn, *Tetraodon nigroviridis*; Ol, *Oryzias latipes*; Ga, *Gasterosteus aculeatus*; Lc, *Latimeria chalumnae*; Xt, *Xenopus tropicalis*; Ac, *Anolis carolinensis*; Gg, *Gallus gallus*; Mm, *Mus musculus*; Hs, *Homo sapiens*; Ci, *Ciona intestinalis*.

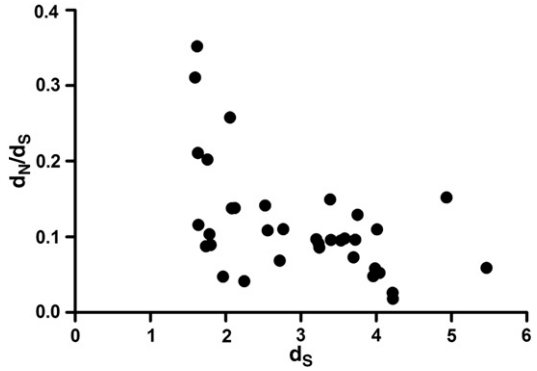


Fig. 3. Individual d_S and d_N/d_S of paralogous START gene pairs in the teleost.

observed. It is possible that extra duplication of this STARD9 gene occurred in the teleost ancestor but was lost before divergence of the zebrafish.

3.3. Divergence and relative rate test of extra-duplicated teleost START pairs

Modes of selection can be estimated by the ratio of the number of nonsynonymous substitutions per nonsynonymous site (d_N) to the number of synonymous substitutions per synonymous site (d_S), i.e. $d_N/d_S>1$ indicates positive selection; $d_N/d_S<1$ indicates purifying selection; and $d_N/d_S=1$ indicates neural evolution (Yang, 2007). The combination of phylogenetic and syntenic analyses revealed extra copies of *COL4A3BP*, *STARD13* and *ACOT11* genes in teleosts. These teleost genes were selected for further evolutionary analysis. Although positive selection has been pervasive during vertebrate evolution (Studer et al., 2008), none of the three site-specific positive selection models or branch site test 2 of PAML predicted any site under positive selection for any teleost START gene listed above with probabilities >95% (data not shown), and pairwise comparison of d_N/d_S between these duplicate pairs was markedly <1 (Fig. 3 Supplemental Table 3), suggesting that these ancient duplicates likely have been subject to purifying selection. An asymmetric molecular evolution rate deviating from the neutral expectation occurs in 7 of 14 fish-specific duplicated teleost START pairs (Table 1), suggesting that one paralog might evolve faster than another after duplication.

3.4. Differential transcript profiling of teleost START genes under experimental conditions

The level at which a gene is expressed under some conditions can provide useful clues to gene function. To examine the transcript abundance patterns of the START genes, we used a comprehensive expression analysis with the publicly available microarray data for zebrafish (Domazet-Loso and Tautz, 2010), medaka (Iwahashi et al., 2009), Green Spotted Puffer (Chan et al., 2009) and three-spined stickleback. All of the 18 three-spined stickleback START genes and 8 of the Green Spotted Puffer START genes were expressed in all detected tissues, but the mRNA level of different genes peaked in different tissues. The difference of steady-state levels of START transcripts between tissues was greater than that between environmental populations. In zebrafish, 14 START genes were detected to be expressed during the ontogenetic progression phase, and expression profiles of them defined clearly different developmental stages (Fig. 4). Differences in transcript abundance levels of these START genes, such as the 400-fold difference of mean expression level between *stard10-3* and *stard10-2* during these stages, were observed. It indicated that the contributions of different STARTs to growth and development might be associated with their expression levels. According to the transcript profiling, 3 zebrafish START gene clusters were observed and the 8 START groups defined in our earlier phylogenetic analysis were all included in these clusters except the data-deficient STARD9 groups, but these clusters were not highly related with gene phylogeny. Contrary expression patterns within members of the STARD1/STAR, STARD4 and STARD10 groups in zebrafish, STARD1/STAR in Green Spotted Puffer and STARD2 in medaka were observed, suggesting divergent functions within the same group during the developmental process.

Transcriptional expression analysis of fish-specific duplicated START paralogs revealed that zebrafish COL4A3BP paralogs and three-spined stickleback ACOT11 paralogs have divergent expression patterns. Of the 3 duplicated zebrafish START10 paralogs identified by our phylogenetic study, *stard10-1* and *stard10-2*, but not *stard10-3*, have similar transcriptional expression patterns (Fig. 4). It appears that the transcriptional expression patterns of the paralogs have diverged during long-term evolution, suggesting functional diversification of duplicated genes.

4. Discussion

As the major organic constituents of fish, lipids function as major sources of metabolic energy for growth, reproduction and migration

Table 1
Tajima relative rate tests of teleost START duplicate genes[a].

Test group	*Mt*[b]	*M1*[c]	*M2*[d]	χ^2	*P*[e]
*col4a3bp-2*_Dr/*col4a3bp-1*_Dr with *COL4A3BP*_Hs	389	68	26	18.77	0.00001
*COL4A3BP-1*_Dr/*COL4A3BP-2*_Tn with *COL4A3BP*_Hs	490	43	48	0.27	0.60018
*COL4A3BP-1*_Ol/*COL4A3BP-2*_Ol with *COL4A3BP*_Hs	495	40	45	0.29	0.58759
*COL4A3BP-1*_Ga/*COL4A3BP-2*_Ga with *COL4A3BP*_Hs	488	31	35	0.24	0.62246
*COL4A3BP-1*_Tr/*COL4A3BP-2*_Tr with *COL4A3BP*_Hs	496	40	44	0.19	0.66252
*stard13-1*_Dr/*stard13-2*_Dr with *STARD13*_Hs	705	55	74	2.80	0.09436
*STARD13-1*_Tn/*STARD13-2*_Tn with *STARD13*_Hs	606	73	149	26.02	0.00000
*STARD13-1*_Ga/*STARD13-2*_Ga with *STARD13*_Hs	616	62	161	43.95	0.00000
*STARD13-1*_Tr/*STARD13-2*_Tr with *STARD13*_Hs	619	71	155	31.22	0.00000
*acot11-1*_Dr/*acot11-2*_Dr with *ACOT11*_Hs	255	25	18	1.14	0.28575
*ACOT11-1*_Tn/*ACOT11-2*_Tn with *ACOT11*_Hs	356	36	66	8.82	0.00297
*ACOT11-1*_Ol/*ACOT11-2*_Ol with *ACOT11*_Hs	362	47	56	0.79	0.37519
*ACOT11-1*_Ga/*ACOT11-2*_Ga with *ACOT11*_Hs	375	41	64	5.04	0.02480
*ACOT11-1*_Tr/*ACOT11-2*_Tr with *ACOT11*_Hs	368	42	70	7.00	0.00815

[a] The Tajima relative rate test was used to examine the equality of evolutionary rate between teleost START duplicate pairs.
[b] *Mt* is the sum of identical sites and divergent sites in all three sequences tested.
[c] *M1* is the number of unique differences in the first paralog.
[d] *M2* is the number of unique differences in the second paralog.
[e] If $P<0.05$ the test rejects the equal substitution rates between the two duplicates.

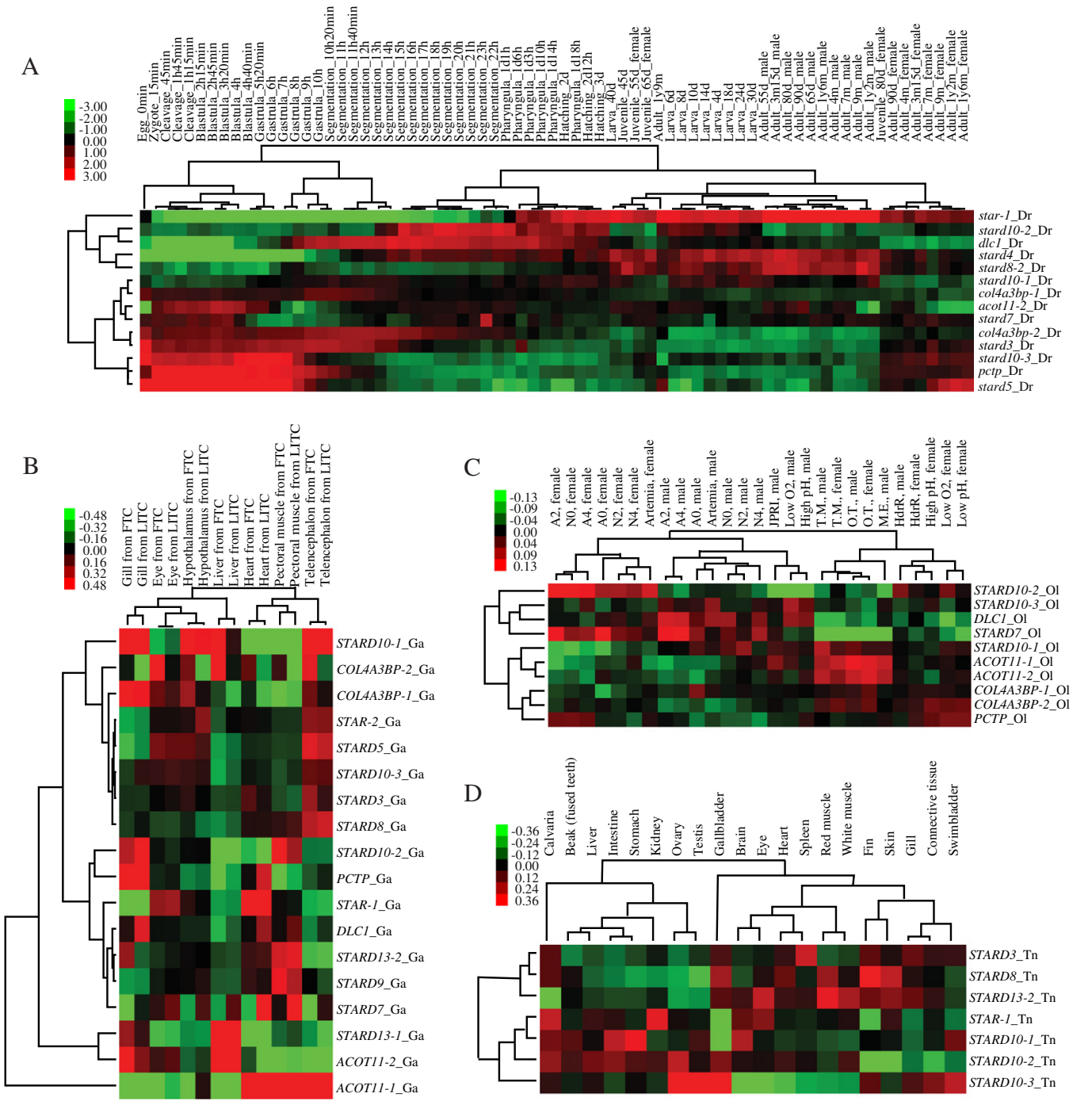


Fig. 4. Relative transcript abundance profiles of the teleost START genes under different conditions. (A) Transcript profiling of the zebrafish START genes (GEO GSE24616) at different developmental stages. (B) Transcriptional expression pattern of three-spined stickleback (GEO GSE12976) START genes in different tissues in marine (LITC) and freshwater (FTC) populations. (C) Transcript abundance pattern of START genes (GEO GSE15380) in different medaka strains (HdrR; JPRI) under different test conditions (A0, aeration and a static water supply; A2, aeration and two times semistatic; A4, aeration and four times semistatic; N0, nonaeration and static; N2, nonaeration and two times semistatic; N4, nonaeration and four times semistatic) or feeding types (Artemia, *Artemia nauplii*; T.M., tetramine; O.T., otohime; M.E., medakanoesa). (D) Transcriptional expression pattern of Green Spotted Puffer (GEO GSE34783) STARTs in different tissues. The transcript abundance levels for the teleost START genes were clustered using hierarchical clustering based on Euclidean distance with complete linkage in the Cluster program. Each row corresponds to the normalized expression profile of a particular gene and their names are shown. The relative transcript accumulation is represented in a color code with green or red showing the lower or higher levels of transcriptional expression, respectively.

(Tocher, 2003). In addition, the fatty acids of fish lipids are rich in ω3 long chain, highly unsaturated fatty acids that have particularly important roles in animal nutrition. However, the metabolic processes regulating deposition and mobilization of fat in fish species are poorly understood (Mommsen et al., 1999). Lipid transfer reactions and START genes have a major role in lipid metabolism, and its disorder is potentially linked to some cardiovascular diseases in human (Tall et al., 1986). However, START genes and their physiological functions in fish are relatively unknown. Here, we present a comparative genomic study of START paralogs in the teleost lineage, and asymmetric evolution and

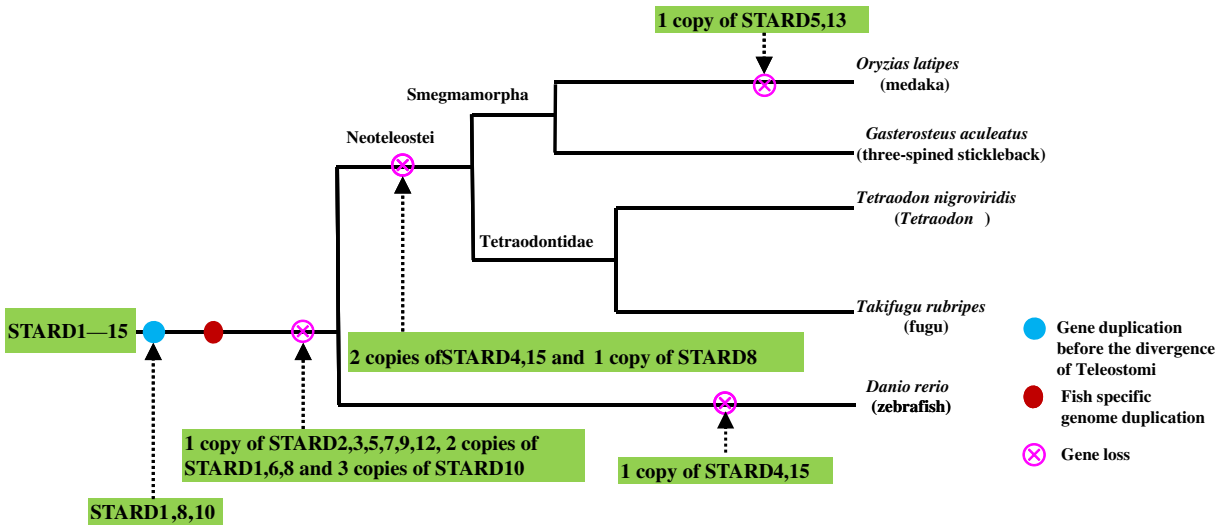


Fig. 5. Hypothetical scenarios of teleost START evolution. The inferred evolutionary events (gene duplication and gene loss) are indicated on the respective branches.

distinct transcriptional expression of these genes in teleosts have been observed. According to our phylogenetic and syntenic analyses, we provide a hypothetical scenario of teleost START evolution (Fig. 5). It is likely that before the divergence of tetrapods and teleosts, 3 (STARD1/8/10) of the 15 ancestral STARTs gave rise to extra duplication followed by mammalian-specific gene losses. After that, extra paralogs of STARTs were generated in the teleost lineage during the FSGD event, followed by the loss of many copies of START genes. After the divergence of zebrafish and Neoteleostei, STARD4/15 and one copy of STARD8 were lost from the Neoteleostei. During successive divergence of the Tetraodontidae and Smegmamorpha and speciation, one copy of STARD5/13 was lost in medaka. This led to the preservation of different ancient STARTs in different teleost fish species.

Eight not 6 START groups were found in this study, because more START genes were used in this study compared to earlier reports (Alpy and Tomasetto, 2005; Soccio and Breslow, 2003). Asymmetry of paralogous gene retention was found between or within each teleost START group. For example, in the STARD1/STAR group, all STARD1 but not all STARD3 were retained in the teleost genome; all teleost STARD6 but partially STARD4 lost in the STARD4 group. Members within the same group might have similar functions. This was confirmed by the observation that (1) the STARD1/STAR group (STARD1 and STARD3) has similar biophysical and functional properties (Tuckey et al., 2004) and mice lacking the STARD3 appear normal and show no defect in steroidogenesis (Kishida et al., 2004); and (2) expression of STARD4 or STARD5 stimulates steroidogenesis by P450scc and liver X receptor reporter gene activity, indicating that both of them function in cholesterol metabolism (Soccio et al., 2005). It is possible that the lost genes were functionally unimportant or redundant to the teleost, or compensation within the same group was available. Intriguingly, the clusters of gene/transcript expression profiles define clearly different developmental stages or environmental conditions, whereas they are not highly related to gene phylogeny. The transcriptional expression differences within members of the STARD1/STAR group were observed in different tissues of Green Spotted Puffer and three-spined stickleback. Earlier studies demonstrated that STARD1 and STARD3 are differentially localized in cells (Alpy et al., 2001) and STARD3 can function in steroidogenesis in organs that do not express STARD1, such as the placenta (Watari et al., 1997). It is possible that tissue-specific expression and subcellular localizations within the same group lead to the observed expression difference during development, because whole fertilized eggs, embryos or larvae were used in the expression study (Domazet-Loso and Tautz, 2010).

It is worthy to mention that 3 teleost STARD10 paralogs were found in our phylogenetic study. Two STARD10 genes that reside in the same chromosome were found in 3 of the 5 teleost fishes, with the exception of incompletely assembled fugu scaffolds or medaka ultracontig. We speculate that tandem duplication occurs in this gene. Because of frequent gene-linkage disruptions, micro-inversions or rearrangements in teleosts, we cannot find further evidence to support the hypothesis of the occurrence of tandem duplication during the evolutionary history of this gene. The expression pattern of duplicated genes can provide useful clues to gene function, and will be of benefit to understanding the driving force and the functional consequence of paralogs (Prince and Pickett, 2002). Zebrafish START10 paralogs have inconsistent transcriptional expression patterns during developmental phases. The transcript abundance of *STARD10-2* and *STARD10-3* peaked in the ovary and testis of Green Spotted Puffer. Diversified STARD10 paralogs might have a role in energy metabolism by mobilizing phosphatidylcholine during development in the testis (Yamanaka et al., 2000). We suspect that the expression difference increases the adaptability of duplicated genes to environmental changes, thus conferring a possible evolutionary advantage.

Asymmetric evolution might be an indicator of neo-functionalization. Some duplicated genes exhibiting asymmetric protein sequence evolution have been reported (Brunet et al., 2006; Conant and Wagner, 2003; Jordan et al., 2004; Lynch and Force, 2000; Nembaware et al., 2002; Van de Peer et al., 2001). This asymmetry has been regarded as a contribution to Ohno's model (Kellis et al., 2004), which proposes that the slow copy maintains an ancestral role and rate of change; while the fast copy evolves to optimize novel functions (Ohno, 1970). Our study revealed fish-specific duplicated extra copies of *col4a3bp*, *stard13* and *acot11* genes in teleosts. When these genes were selected for further evolutionary analysis, we found that an asymmetric molecular evolution rate deviating from the neutral expectation occurs in 7 of 14 duplicated pairs (Table 1). Similar to our analysis, the duplicated teleost HoxA clusters or type III receptor tyrosine kinase (RTK) genes were characterized as evolvement in an asymmetric manner (Braasch et al., 2006; Wagner et al., 2005). These results indicate that asymmetric divergence of fish-specific paralogs might be a common feature, and this feature might contribute to some fish-specific behavior or the diversity of teleost fishes.

In conclusion, asymmetric evolution and divergent transcriptional expression of START genes have occurred in teleost genomes. This detailed analysis of the START gene family in teleost fishes has provided a case in studying how genes evolve after gene duplication, and might

provide some insights into the physiological function divergence of START genes in fishes.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2013.01.058.

Acknowledgments

This work was supported by the grants from the National Natural Science Foundation of China (31200888), the International S&T Cooperation Program of China (2011DFA30670) and the Wenzhou Science and Technology Program (S20100054).

References

Alpy, F., Tomasetto, C., 2005. Give lipids a START: the StAR-related lipid transfer (START) domain in mammals. J. Cell Sci. 118, 2791–2801.
Alpy, F., et al., 2001. The steroidogenic acute regulatory protein homolog MLN64, a late endosomal cholesterol-binding protein. J. Biol. Chem. 276, 4261–4269.
Amores, A., et al., 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711–1714.
Aparicio, S., et al., 2002. Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. Science 297, 1301–1310.
Braasch, I., Salzburger, W., Meyer, A., 2006. Asymmetric evolution in two fish-specifically duplicated receptor tyrosine kinase paralogons involved in teleost coloration. Mol. Biol. Evol. 23, 1192–1202.
Brunet, F.G., et al., 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol. 23, 1808–1816.
Chan, E.T., et al., 2009. Conservation of core gene expression in vertebrate tissues. J. Biol. 8, 33.
Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S., Venkatesh, B., 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol. Biol. Evol. 21, 1146–1151.
Conant, G.C., Wagner, A., 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. 13, 2052–2058.
Conant, G.C., Wolfe, K.H., 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9, 938–950.
de Hoon, M.J., Imoto, S., Nolan, J., Miyano, S., 2004. Open source clustering software. Bioinformatics 20, 1453–1454.
Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N., Hahn, M.W., 2006. The evolution of mammalian gene families. PLoS One 1, e85.
Domazet-Loso, T., Tautz, D., 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468, 815–818.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Flicek, P., et al., 2011. Ensembl 2011. Nucleic Acids Res. 39, D800–D806.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531–1545.
Haldane, J.B.S., 1932. The Causes of Evolution. Harper & Brothers, New York, London.
Hanada, K., Kuromori, T., Myouga, F., Toyoda, T., Shinozaki, K., 2009. Increased expression and protein divergence in duplicate genes is associated with morphological diversification. PLoS Genet. 5, e1000781.
He, X., Zhang, J., 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169, 1157–1164.
Iwahashi, H., Kishi, K., Kitagawa, E., Suzuki, K., Hayashi, Y., 2009. Evaluation of the physiology of Medaka as a model animal for standardized toxicity tests of chemicals by using mRNA expression profiling. Environ. Sci. Technol. 43, 3913–3918.
Iyer, L.M., Koonin, E.V., Aravind, L., 2001. Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. Proteins 43, 134–144.
Jaillon, O., et al., 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946–957.
Johnson, L.S., Eddy, S.R., Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinforma. 11, 431.
Jordan, I.K., Wolf, Y.I., Koonin, E.V., 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol. Biol. 4, 22.
Kasahara, M., et al., 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447, 714–719.
Kellis, M., Birren, B.W., Lander, E.S., 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428, 617–624.
Kishida, T., et al., 2004. Targeted mutation of the MLN64 START domain causes only modest alterations in cellular sterol metabolism. J. Biol. Chem. 279, 19276–19285.
Lynch, M., Force, A., 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154, 459–473.
Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T., 2010. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 39, D52–D57.
Mistry, J., Finn, R., 2007. Pfam: a domain-centric method for analyzing proteins and proteomes. Methods Mol. Biol. 396, 43–58.
Mommsen, T.P., Vijayan, M.M., Moon, T.W., 1999. Cortisol in teleosts: dynamics, mechanisms of action, and metabolic regulation. Rev. Fish Biol. Fish. 9, 211–268.
Nembaware, V., Crum, K., Kelso, J., Seoighe, C., 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res. 12, 1370–1376.
Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, New York.
Ponting, C.P., Aravind, L., 1999. START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. Trends Biochem. Sci. 24, 130–132.
Postlethwait, J.H., 2007. The zebrafish genome in context: ohnologs gone missing. J. Exp. Zool. B Mol. Dev. Evol. 308, 563–577.
Prince, V.E., Pickett, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. 3, 827–837.
Sato, Y., Hashiguchi, Y., Nishida, M., 2009. Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. BMC Evol. Biol. 9, 127.
Schrick, K., Nguyen, D., Karlowski, W.M., Mayer, K.F., 2004. START lipid/sterol-binding domains are amplified in plants and are predominantly associated with homeodomain transcription factors. Genome Biol. 5, R41.
Soccio, R.E., Breslow, J.L., 2003. StAR-related lipid transfer (START) proteins: mediators of intracellular lipid metabolism. J. Biol. Chem. 278, 22183–22186.
Soccio, R.E., Adams, R.M., Maxwell, K.N., Breslow, J.L., 2005. Differential gene regulation of StarD4 and StarD5 cholesterol transfer proteins. Activation of StarD4 by sterol regulatory element-binding protein-2 and StarD5 by endoplasmic reticulum stress. J. Biol. Chem. 280, 19410–19418.
Studer, R.A., Penel, S., Duret, L., Robinson-Rechavi, M., 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. Genome Res. 18, 1393–1402.
Tall, A., Sammett, D., Granot, E., 1986. Mechanisms of enhanced cholesteryl ester transfer from high density lipoproteins to apolipoprotein B-containing lipoproteins during alimentary lipemia. J. Clin. Invest. 77, 1163–1172.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Taylor, J.S., Raes, J., 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. 38, 615–643.
Teng, H., et al., 2010. Evolutionary mode and functional divergence of vertebrate NMDA receptor subunit 2 genes. PLoS One 5, e13342.
Tocher, D.R., 2003. Metabolism and functions of lipids and fatty acids in teleost fish. Rev. Fish. Sci. 11, 107–184.
Tuckey, R.C., Bose, H.S., Czerwionka, I., Miller, W.L., 2004. Molten globule structure and steroidogenic activity of N-218 MLN64 in human placental mitochondria. Endocrinology 145, 1700–1707.
Van de Peer, Y., Taylor, J.S., Braasch, I., Meyer, A., 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J. Mol. Evol. 53, 436–446.
Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A., Van de Peer, Y., 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc. Natl. Acad. Sci. U. S. A. 101, 1638–1643.
Wagner, G.P., et al., 2005. Molecular evolution of duplicated ray finned fish HoxA clusters: increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. J. Mol. Evol. 60, 665–676.
Watari, H., et al., 1997. MLN64 contains a domain with homology to the steroidogenic acute regulatory protein (StAR) that stimulates steroidogenesis. Proc. Natl. Acad. Sci. U. S. A. 94, 8462–8467.
Woods, I.G., et al., 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res. 15, 1307–1314.
Yamanaka, M., et al., 2000. Molecular cloning and characterization of phosphatidylcholine transfer protein-like protein gene expressed in murine haploid germ cells. Biol. Reprod. 62, 1694–1701.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.