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

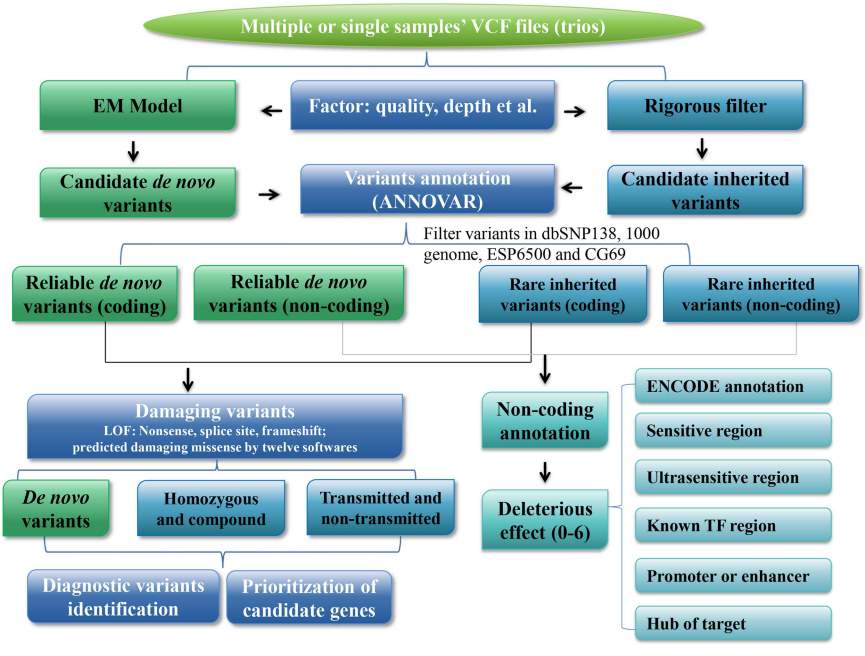


Figure 1 The workflow of mirTrios. mirTrios embarks on the analysis of multiple or single trios-based variant call files (VCF). The workflow and results of mirTrios comprise flowing parts: (1) detection of de novo mutations (DNMs) based on expectation-maximisation (EM) modules; (2) detection of inherited 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).

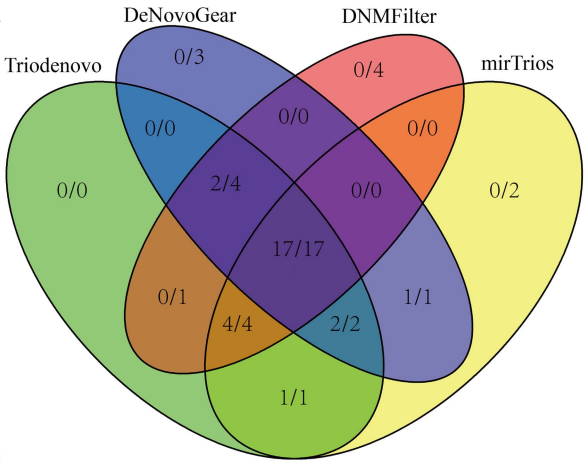
RESULTS

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

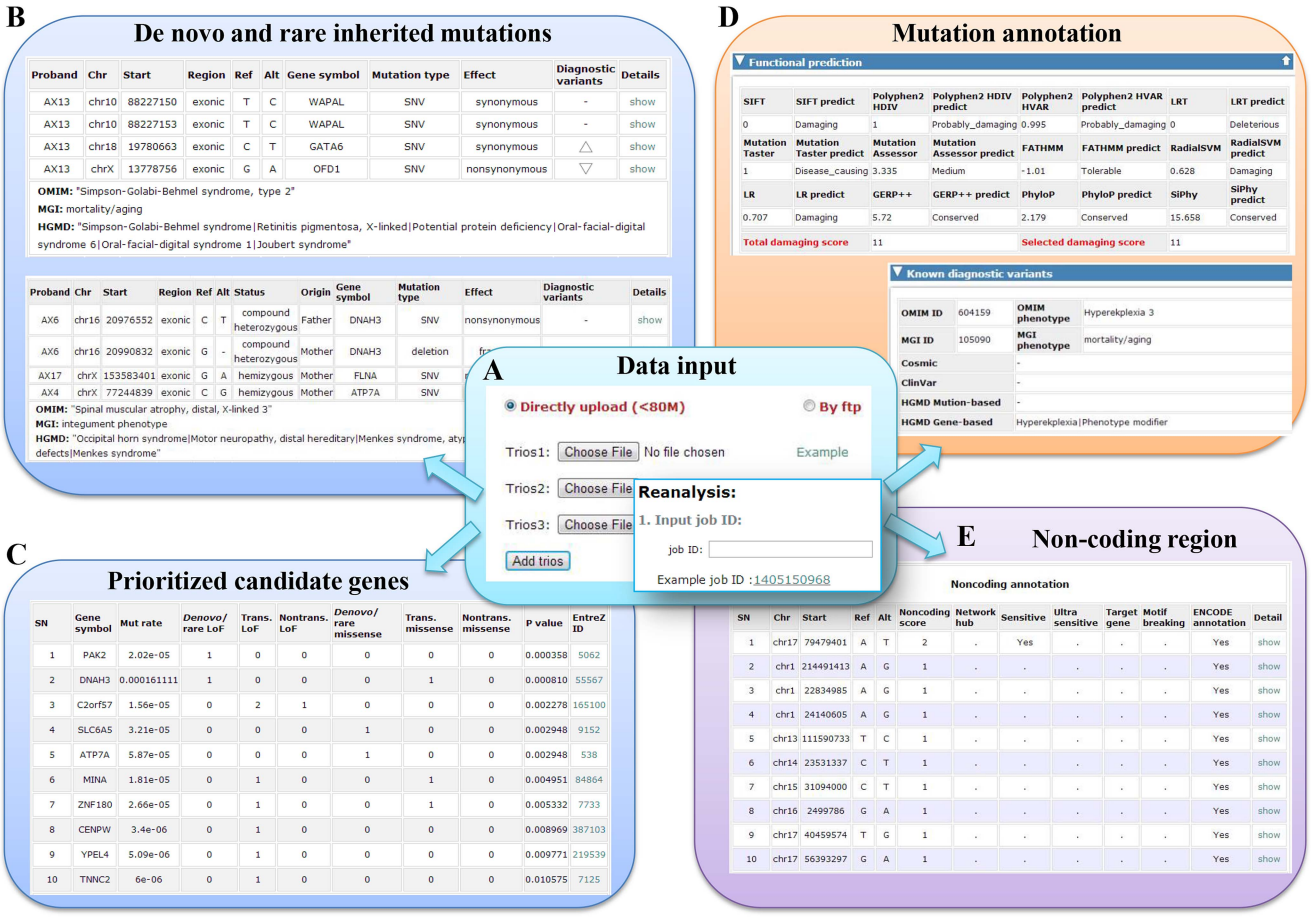


Figure 3 The snapshot of the results of mirTrios. (A) Trios-based variant call files (VCF) and family list are loaded as input along with several user-selected options. (B) Detected de novo mutations (DNMs) and rare inherited mutations. (C) Comprehensive annotation of detected variations. (D) Known diagnostic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