Supplementary materials

Supplement to: Intraventricular fetus-in-fetu with extensive de novo gain in genetic copy number

Online Method

Genetic sequencing and data analysis

DNA extract and detect

Genomic DNA extracted from the tissue of the fetus-in-fetu and the peripheral blood of the host child and parents was fragmented to an average size of ~350bp and subjected to DNA library creation using established Illumina paired-end protocols. The Illumina Novaseq 6000 platform (Illumina Inc., San Diego, CA, USA) was utilized for genomic DNA sequencing in Novogene Bioinformatics Technology Co., Ltd (Beijing, China) to generate 150-bp paired-end reads with a minimum coverage of 10× for 99% of the genome (mean coverage of 30×).

Data analysis

After sequencing, basecall files conversion and demultiplexing were performed with bcl2fastq software (Illumina). The resulting fastq data were submitted to in-house quality control software for removing low quality reads, and then were aligned to the reference human genome (hs37d5) using the Burrows-Wheeler Aligner (bwa)[1], and duplicate reads were marked using sambamba tools[2].

SNP/INDEL calling

Single nucleotide variants (SNVs) and indels were called with samtools to generate gVCF[3]. The raw calls of SNVs and INDELs were further filtered with the following inclusion thresholds: 1) read depth > 4; 2) Root-Mean-Square mapping quality of covering reads > 30; 3) the variant quality score > 20.

CNV calling

The copy number variants (CNVs) were detected with software Control-FREEC(v9.1) [4], using a 1-kb as threshold of duplication and deletion.

SV calling

The structural variants(SVs) were detected with software LUMPY(version v0.2.13)[5].

Annotation

Annotation was performed using ANNOVAR (2017June8) [6]. Annotations included minor allele frequencies from public control data sets as well as deleteriousness and conservation scores enabling further filtering and assessment of the likely pathogenicity of variants.

Rare variants filtering

Filtering of rare variants was performed as follows: (1) variants with a MAF less than 0.01 in 1000 genomic data (1000g_all)[7], esp6500siv2_all[8], gnomAD data (gnomAD_ALL and gnomAD_EAS)[9] and in house Novo-Zhonghua exome database

from Novogene; (2) Only SNVs occurring in exons or splice sites (splicing junction 10 bp) are further analyzed since we are interested in amino acid changes. (3) Then synonymous SNVs which are not relevant to the amino acid alternation predicted by dbscSNV are discarded; The small fragment non-frameshift (<10bp) indel in the repeat region defined by RepeatMasker are discarded. (4) Variations are screened according to scores of SIFT[10], Polyphen[11], MutationTaster[12] and CADD[13] softwares. The potentially deleterious variations are reserved if the score of more than half of these four softwares support harmfulness of variations[14]. Sites(>2bp) did not affect alternative splicing were removed.

4. Kinship analysis

Relationship between proband and parents was estimated using the pairwise identity-by-descent (IBD) calculation in PLINK[15]. The IBD sharing between the proband and parents in all trios is over 50%.

References

[1] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25(14), 1754 (2009).

[2] A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015.

[3] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools[J]. *Bioinformatics*. 2009, 25(16): 2078-2079.

[4] Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data[J]. *Bioinformatics*. 2012, 28(3): 423-425.

[5] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84.

[6] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010, 38(16): e164-e164.

[7] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015 Oct 1; 526(7571):68–74.

[8] Exome Variant Server, NHLB GO Exome Sequencing Project (ESP), Seattle, WA. Available at: http://evs.gs.washington.edu/EVS. Accessed February/21, 2017.

[9] gnomAD. Available at: https://doi.org/10.1101/030338. Accessed February/21, 2017.

[10] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009; 4:1073–1081.

[11] Adzhubei IA, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010; 7:248–249.

[12] Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods*. 2010; 7:575–576.

[13] Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014; 46:310–315.

[14] Muona, M., Berkovic, S. F., Dibbens, L. M., Oliver, K. L., Maljevic, S., Bayly, M. A., et al. A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nature Genetics*. 47(1), 39–46.

[15] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based

linkage analyses. *Am J Hum Genet*. 2007;81(3):559-75.