


Whole genome sequencing of 10K patients with acute ischaemic stroke or transient ischaemic attack: design, methods and baseline patient characteristics

Si Cheng,^{1,2,3} Zhe Xu,^{1,2,3} Yang Liu,^{1,2,3} Jinxi Lin,^{1,2} Yong Jiang,^{1,2} Yilong Wang,^{1,2} Xia Meng,^{1,2} Anxin Wang ,^{1,2} Xinying Huang,^{1,2} Zhimin Wang,⁴ Guohua Chen,⁵ Songdi Wu,⁶ Zhengchang Jia,⁷ Yongming Chen,⁸ Xuerong Qiu,⁹ Jun Wu,¹⁰ Binbin Song,¹¹ Weizhong Ji,¹² Zhongping An,¹³ Wenjun Xue,¹⁴ Lili Zhao,¹⁵ Yu Geng,¹⁶ Hongyan Li,¹⁷ Hao Li ,^{1,2} Yongjun Wang ^{1,2,3}

To cite: Cheng S, Xu Z, Liu Y, *et al.* Whole genome sequencing of 10K patients with acute ischaemic stroke or transient ischaemic attack: design, methods and baseline patient characteristics. *Stroke & Vascular Neurology* 2021;**6**: e000664. doi:10.1136/svn-2020-000664

SC, ZX and YL are joint first authors.

Received 4 October 2020
Revised 29 October 2020
Accepted 17 November 2020
Published Online First
18 December 2020



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Yongjun Wang;
yongjunwang@nrcrnd.org.cn

ABSTRACT

Background and purpose Stroke is the second leading cause of death worldwide and the leading cause of mortality and long-term disability in China, but its underlying risk genes and pathways are far from being comprehensively understood. We here describe the design and methods of whole genome sequencing (WGS) for 10 914 patients with acute ischaemic stroke or transient ischaemic attack from the Third China National Stroke Registry (CNSR-III).

Methods Baseline clinical characteristics of the included patients in this study were reported. DNA was extracted from white blood cells of participants. Libraries are constructed using qualified DNA, and WGS is conducted on BGISEQ-500 platform. The average depth is intended to be greater than 30× for each subject. Afterwards, Sentieon software is applied to process the sequencing data under the Genome Analysis Toolkit best practice guidance to call genotypes of single nucleotide variants (SNVs) and insertion-deletions. For each included subject, 21 fingerprint SNVs are genotyped by MassARRAY assays to verify that DNA sample and sequencing data originate from the same individual. The copy number variations and structural variations are also called for each patient. All of the genetic variants are annotated and predicted by bioinformatics software or by reviewing public databases.

Results The average age of the included 10 914 patients was 62.2±11.3 years, and 31.4% patients were women. Most of the baseline clinical characteristics of the 10 914 and the excluded patients were balanced.

Conclusions The WGS data together with abundant clinical and imaging data of CNSR-III could provide opportunity to elucidate the molecular mechanisms and discover novel therapeutic targets for stroke.

INTRODUCTION

Stroke is the second leading cause of death worldwide, and the leading cause of mortality and long-term disability in China.¹

Being the most common type of stroke, ischaemic stroke (IS) accounts for about 80% of all strokes,² and more than 90% of IS are sporadic.³ IS is a complex multifactorial disease arising from complicated gene-environment interactions. Therefore, uncovering genetic contributions to IS could help to identify the genes, pathways and networks that are involved in IS pathogenesis. Although several novel genetic variants that were associated with IS susceptibility have been discovered in the last decades,^{4–9} few studies explored the correlation between genetic variants and stroke outcomes. Moreover, previous genetic studies on IS were mainly conducted in European and African populations,^{4 10} and there is limited data for the Chinese population. Due to the substantial ancestral difference,¹¹ whether these reported IS-associated genetic variants could also contribute to IS pathogenesis in Chinese population needs verification.

The Third China National Stroke Registry (CNSR-III) is a nationwide prospective registry with 15 166 patients with IS or transient ischaemic attack (TIA) in China.¹² A broad and comprehensive spectrum of individual-level data had been collected, including clinical phenotypes, aetiological classification, neuroimaging, biomarkers and clinical outcomes. The aetiological subtyping information was recorded centrally. Taking these advantages, we perform whole genome sequencing (WGS) for 10 914 patients in the prespecified genetic substudy of CNSR-III to delineate the genetic landscape of IS and TIA in Chinese population.

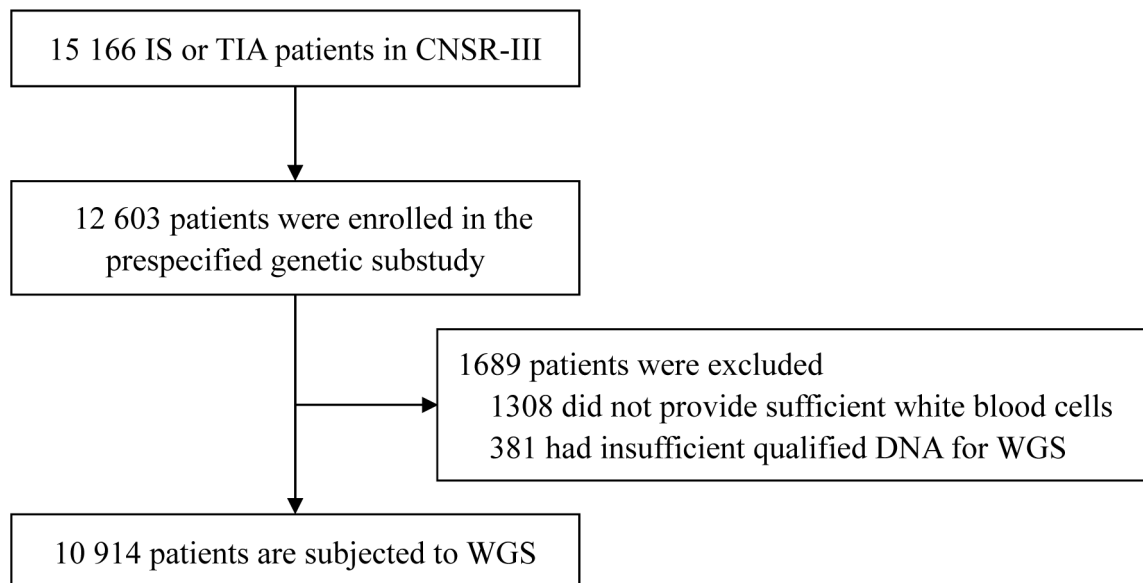


Figure 1 Flow chart of patient selection for WGS in the prespecified genetic substudy of CNSR-III. CNSR-III, The Third China National Stroke Registry; IS, ischaemic stroke; TIA, transient ischaemic attack; WGS, whole genome sequencing.

METHODS

Patients

The CNSR-III is a nationwide prospective registry for patients presented to hospitals with acute ischaemic cerebrovascular events between August 2015 and March 2018 in China.¹² There is a total of 15 166 patients with IS (n=14 146, 93.3%) or TIA (n=1020, 6.7%) within 7 days from the onset of symptoms to enrollment. The CNSR-III involved 201 hospitals that cover 22 provinces and 4 municipalities in China, including 163 grade III (central hospitals for certain district or city, usually teaching hospitals) and 38 grade II (hospitals serving several communities) urban hospitals. A total of 12 603 patients participated in the prespecified genetic substudy. The white blood cells (WBCs) from a total of 10 914 patients are applied in WGS (figure 1). The written informed consents were obtained from all patients or legally authorised representatives before entering into the study.

DNA extraction

For each sample, WBCs was used to extract the genomic DNA, which was performed using Magnetic Blood Genomic DNA Kit (DP329, TIANGEN Biotech Co Ltd, Beijing, China) on KingFisher Flex (Thermo Scientific Co, Massachusetts, USA) system for automatic genomic DNA extraction and purification at iGeneTech Co Ltd. (Beijing, China) or by manual phenol–chloroform DNA extraction at BGI Genomics (BGI-Shenzhen).

Evaluation of DNA quality

The concentration of genomic DNA was quantified using Qubit 2.0 fluorometer (Thermo Scientific Co, Massachusetts, USA) and SpectraMax Gemini XPS (Molecular Devices, San Francisco, USA) at BGI Genomics (BGI-Shenzhen). Electrophoresis was conducted on 1% agarose gel to make sure that the majority of genomic

DNA segments was longer than 20 Kb and was not substantially degraded. Genomic DNA samples with concentration ≥ 12.5 ng/ μ L and total amount ≥ 0.5 μ g was qualified for further procedures. For each of the qualified sample, the DNA is further applied in library construction and subsequent WGS process, as well as single nucleotide variant (SNV) genotyping (see details below).

Library construction

The qualified genomic DNA is randomly fragmented by ultrasound using CovarisLE220 (Covaris, Massachusetts, USA) according to the manufacturer's instructions. The DNA fragments in the range of 200 to 400 bp are selected by VAHTSTM DNA Clean Beads (Vazyme Biotech Co, Ltd, Nanjing, China). The end repair for DNA fragments is performed by adding an 'A' nucleotide to the 3' end of each strand. Afterwards, the dTTP-tailed adapters are ligated to both ends of the repaired/dA-tailed DNA fragments. The ligation product is then amplified by PCR. Then the products are purified by VAHTSTM DNA Clean Beads (Vazyme Biotech Co, Ltd, Nanjing, China). The purified PCR products with total mass ≥ 200 ng, and the main peak in 300 to 500 bp should be applied. Single strand separation is conducted by heat-denaturing the PCR product at 95 °C. Circularisation process is performed by mixing the single-stranded DNA fragments with splint oligos (sequence: GCCATGTCGTTCTGTGAGCCAAGG) and DNA Rapid Ligase to generate single-stranded DNA circles. The remaining linear molecule is digested with the exonuclease. The enzymatic digestion products are purified by Agencourt AMPure XP medium (Beckman Coulter, Indiana, USA). The single-stranded circle DNA (ssCir DNA) are formatted as the final library. The purified enzymatic digestion products are quantified with Qubit ssDNA Assay Kit (Thermo

Scientific Co, Massachusetts, USA), and the final yield should be ≥ 12 ng.

BGISEQ-500 WGS sequencing

Rolling circle amplification is performed for the qualified libraries to produce DNA Nanoballs (DNBs). Then the DNBs are loaded into the patterned nanoarrays and 100bp pair-end reads are sequenced on the BGISEQ-500 platform (BGI Genomics, Shenzhen, China). Sequencing-derived raw image files are processed by BGISEQ-500 base-calling software (V.1.2.1.21840) under default parameters settings. The sequence data are stored in FASTQ format. The average depth for each subject is intended to be greater than $30\times$.

SNV genotyping

To make sure that the DNA samples are neither mistaken nor contaminated during the WGS process, we selected 21 biallelic fingerprint SNVs and planned to genotype them for each participant of WGS. These 21 SNVs distribute on 15 different autosomes and are at least 13M apart. The minor allele frequencies of these SNVs are between 0.16 to 0.5 within the Han Chinese in Beijing samples in 1000 Genome Project.¹³ The SNV genotyping experiments are performed at BGI Genomics (BGI-Shenzhen) independently and simultaneously with WGS. For each sample, approximately 30 ng of qualified genomic DNA is used. Locus-specific PCR and detection primers are designed using the MassARRAY Assay Design software (Agena Bioscience, California, USA). Multiplex PCR and locus-specific single-nucleotide extension are performed for each DNA sample, then the products are desalted and transferred to a 384-well SpectroCHIP array. After MALDI-TOF (matrix-assisted laser desorption/ionization-time of flight) mass spectrometry, MassArray Typer software (V.4.1, Agena Bioscience, California, USA) is used to call the genotype for each participant.

After the accomplishment of WGS and SNV genotyping, the genotypes of the 21 SNVs are compared between those that are respectively obtained from WGS data analyses and MALDI-TOF mass spectrometry to verify that DNA sample and sequencing data originates from the same individual.

WGS data cleanup

Raw sequence reads are filtered using an in-house pipeline for quality control. The following steps are executed consecutively: Removing both of the paired reads if (1) any one of the reads contain sequencing adapter, (2) any one of the reads whose low-quality base ratio (base quality less than or equal to 12) is more than 50%, (3) any one of the reads whose unknown base ('N' base) ratio is more than 10%. Afterwards, fastp (V.0.20.0) is applied to filter out low-quality reads and bases,¹⁴ and downstream bioinformatics analyses are conducted on these qualified data.

Mapping and variant calling

The paired-end reads are processed under the Genome Analysis Toolkit (GATK) best practice guidance using

Sentieon (release 201808.05, <https://www.sentieon.com>, bioRxiv 115717; doi:10.1101/115717).¹⁵ The reads are aligned to the hg38 human reference genome sequence that is downloaded from GATK bundle (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Homo_sapiens_assembly38.fasta.gz) using Burrows-Wheeler Alignment tool that is implemented in Sentieon. The SNVs and insertion-deletions (indels) in the regions of segmental duplications and unassigned chromosomes are ignored in the downstream analyses. For each sample, the base quality, sequencing depth, GC (guanine-cytosine) content, mapping rate, mismatch rate, duplication rate and coverage is calculated. After removing the duplicated reads and recalibrating the base quality scores, SNVs and indels are first called using Haplotyper of Sentieon for each individual and then jointly called for all of the participants. Then, variant quality score recalibration and hard filter methods are applied to obtain the high-quality variant calls for SNVs and indels. The '*.bam' and '*.vcf' files that are generated in the above procedures would be reserved for other researches. Copy number variations (CNVs) and structural variations (SVs) in the genome of patients are mainly called using GraphTyper2 and Manta.^{16 17}

Population genetics analysis

To minimise problems arising from hidden family and population structure in the participants, we conduct the following quality control steps. First, kinship is explored by calculating pairwise identity-by-descent calculations for all pairs of individuals using PLINK (V.1.9).¹⁸ The existence of first and second degree relationships is checked using KING (V.2.1.8).¹⁹ Second, population structure is investigated using STRUCTURE software and by conducting principal component analysis.²⁰ All of these analyses are conducted using autosomal SNVs and indels.

Variant annotation

Impact of the mutations on protein coding and protein truncating variants were predicted using variant effect predictor.²¹ Pathogenicity of SNVs and indels are evaluated using InterVar software (V.2.0.1) under guidelines of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.²² The potential impact of SNVs and indels on gene expression/regulation is investigated by reviewing GTEx, HaploReg and other databases or online tools.^{23 24} The impact of intronic and exonic mutations on pre-messenger RNA splicing is mainly predicted using SpliceAI.²⁵

Biological significance of known or common CNVs and SVs are annotated by reviewing dbVar and Database of Genomic Variants.^{26 27} Novel CNVs and SVs are annotated by reviewing literatures on structure and function studies of the genes affected by the corresponding CNVs and SVs from PubMed.

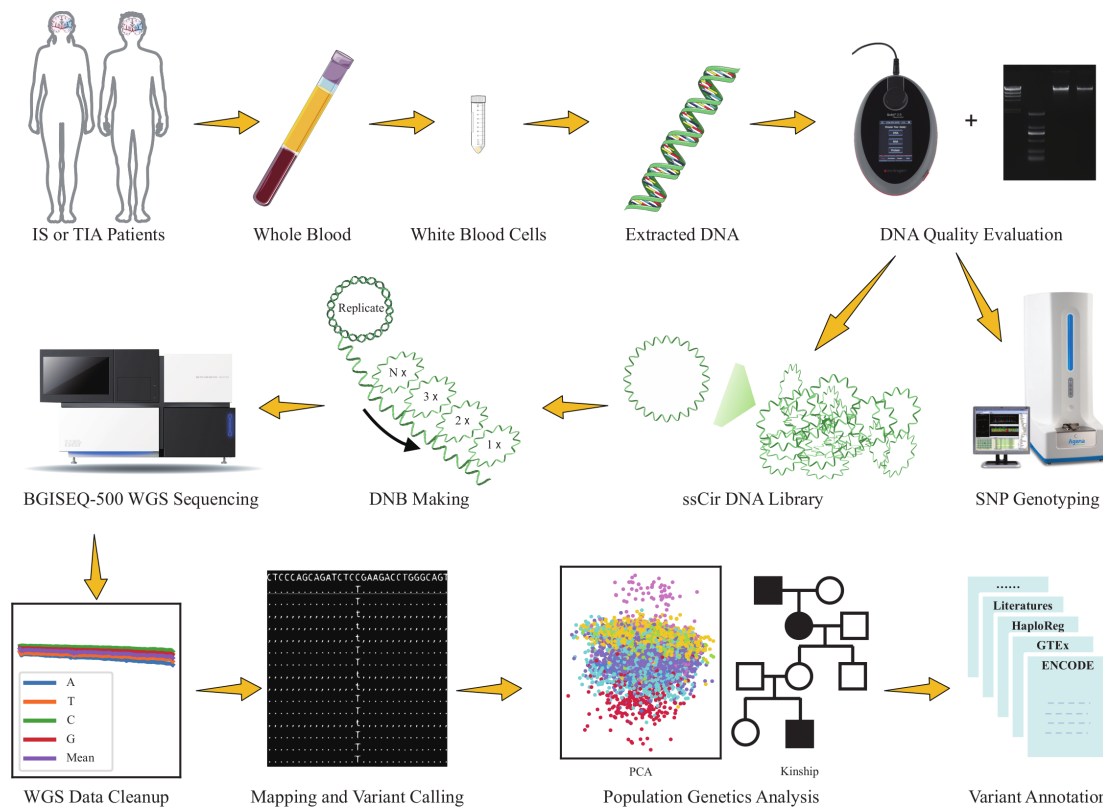


Figure 2 Workflow of WGS and bioinformatics analyses. The first two rows shows the process of DNA extraction, quality control, library construction and WGS. The third row demonstrates downstream bioinformatics analyses of WGS data. Some of the images are retrieved or adapted from Servier Medical Art (<https://smart.servier.com/>), which is licensed under a Creative Commons Attribution 3.0 Unported License. The photos of instruments are downloaded from websites of BGI Genomics (<https://www.bgi.com/>), Thermo Fisher (<https://www.thermofisher.com/>) and Agena Bioscience (<https://agenabio.com/>), respectively. DNB, DNA Nanoball; IS, ischaemic stroke; ssCir DNA, single-stranded circle DNA; TIA, transient ischaemic attack; WGS, whole genome sequencing.

Checking and reviewing

During the experimental procedures of this project, all of the WBCs and DNA loading, packaging, transferring and storing operations was conducted by one technician while being checked and supervised by another technician.

For WGS and SNV genotyping data, the MD5 code is generated for each data file before transfer, and is checked after the transfer. The commands and codes for WGS data mapping and variant calling are written by one bioinformatician while being reviewed by another bioinformatician. The log files are also reviewed and reserved.

All of the genetic information, clinical data and biospecimens are managed following the Regulations of the People's Republic of China on Administration of Human Genetic Resources 2019.

Research projects

WGS data of 10K patients will be incorporated to identify the causality of certain risk factors for stroke outcomes, to investigate pleiotropic effect of genes on multiple phenotypes, and to understand the genetic relationship between particular comorbidities and IS. The accurate sequencing data from greater than 30× average depth in the WGS study also allows us to obtain a panoramic view of individual-specific variation and genetic structure

of Chinese patients with IS or TIA. Some prespecified research topics are described below:

- ▶ To draw a comprehensive genetic landscape of Chinese patients with IS or TIA, and characterise the geographical, lifestyle differences and their demographic origin;
- ▶ To evaluate the genetic contribution to IS and its recurrent outcomes, especially the contribution of rare variants, CNVs and variants in certain region of the genome (eg, telomere and mitochondrial DNA);
- ▶ To determine the causality of serum biomarkers for IS outcomes using association analyses and Mendelian randomisation;
- ▶ To investigate the relationship between genetic features and brain imaging changes in IS;
- ▶ To conduct the pharmacogenomics analyses on certain secondary prevention of IS;
- ▶ To better understand the genetic mechanisms of IS with particular comorbidities (eg, chronic kidney disease, diabetes mellitus and hypertension).

RESULTS

Among the 15 166 patients with IS or TIA in CNSR-III, 12 603 patients participated in the prespecified genetic

Table 1 Baseline characteristics of the included patients in the patients who underwent whole genome sequencing and the rest of the patients in CNSR-III

Characteristics	Included (n=10914 to 72.0%)	Excluded (n=4252 to 28.0%)	Total (n=15 166 to 100%)
Age (years), mean±SD	62.2±11.3	62.2±11.3	62.2±11.3
Female, n (%)	3429 (31.4)	1373 (32.3)	4802 (31.7)
Ethnicity (non-Han), n (%)	306 (2.8)	134 (3.2)	440 (2.9)
Stroke type			
IS	10 166 (93.2)	3980 (93.6)	14 146 (93.3)
TIA	748 (6.8)	272 (6.4)	1020 (6.7)
Current smoker, n (%)	3472 (31.8)	1280 (30.1)	4752 (31.3)
Heavy drinker, n (%)*	1582 (14.5)	544 (12.8)	2126 (14.0)
Medical history, n (%)			
Ischaemic stroke	2322 (21.3)	827 (19.4)	3149 (20.8)
Coronary heart disease	1179 (10.8)	429 (10.1)	1608 (10.6)
Atrial fibrillation	765 (7.0)	254 (6.0)	1019 (6.7)
Hypertension	6858 (62.8)	2636 (62.0)	9494 (62.6)
Diabetes mellitus	2609 (23.9)	901 (21.2)	3510 (23.1)
Hypercholesterolaemia	903 (8.3)	288 (6.8)	1191 (7.9)
NIHSS at admission, median (IQR)†	3.0 (1.0 to 6.0)	3.0 (1.0 to 5.0)	3.0 (1.0 to 6.0)
NIHSS 0–3	5120 (50.4)	2199 (55.2)	7319 (51.7)
NIHSS ≥4	5046 (49.6)	1781 (44.8)	6827 (48.3)

*Heavy drinker was defined as ≥2 standard alcohol consumption per day.

†NIHSS in this table were summarised among IS patients only.

CNSR-III, the Third China National Stroke Registry; IS, ischaemic stroke; NIHSS, National Institutes of Health Stroke Scale score; TIA, transient ischaemic attack.

substudy. Among them, 1308 participants did not provide enough WBCs. After DNA extraction and quality evaluation, the DNA of 381 participants was insufficient or unqualified. Therefore, a total of 1689 participants were excluded and WGS are conducted for 10 914 participants of CNSR-III (figure 1). The workflow of WGS and downstream bioinformatics analyses are shown in figure 2.

Baseline clinical characteristics of the included 10 914 patients and excluded patients were presented in table 1. The average age was 62.2±11.3 years, and 31.4% of the patients were women. Patients diagnosed to be IS were 10 166 (93.2%), among which 50.4% had minor stroke (NIHSS (National Institutes of Health Stroke Scale score) ≤3). A total of 31.8% of the included patients were current smokers, and 14.5% were heavy drinkers (defined as ≥2 standard alcohol consumption per day). A total of 21.3% of the included patients had a history of IS. A total of 10.8%, 7.0% and 62.8% of the included patients had a history of coronary heart disease, atrial fibrillation and hypertension, respectively. The two groups of included and excluded patients were balanced regarding baseline characteristics (table 1).

DISCUSSION

Stroke is a complex disease that has multiple aetiologies. Genetic and genomic studies among populations from diverse ancestry could refine our understanding on

molecular mechanism of stroke. Therefore, we conduct WGS for 10 914 patients from CNSR-III. The WGS procedures and baseline characteristics of patients are reported in this study. The WGS of CNSR-III constructs a genomic data set that facilitate large scale IS genetic analyses in Chinese population. The CNSR-III collected a comprehensive spectrum of phenotypic information under consistent and standardised criteria, which could increase the power and credibility of the genetic analyses. In addition, all of the patients are followed up for clinical outcomes,¹² and this provides an opportunity for discovery of genetic variants that are associated with patients' outcomes after stroke.

In contrast to DNA microarrays that were mainly used in previous genetic associations on IS,^{4 10} WGS technology applied in this study could provide nearly all of the SNVs and indels, and simultaneously capture genetic information on CNVs and SVs for each patient. Therefore, WGS enables a systematic evaluation of the genetic effect of rare variants (allele frequencies <1% in population) to IS and TIA. As the contribution of the rare variants remains one of the top challenges in stroke genetics, the WGS study would provide a better understanding on IS and TIA pathophysiology.¹⁰ The average depth for WGS is intended to be greater than 30× in this project, because at this depth, both accurate variant calling and cost-effectiveness could be achieved.^{28 29} Moreover, >95% the genome could be covered by at least 10

sequencing reads, and >95% of the heterozygous variation could be accurately identified under this design.³⁰ Therefore, the WGS could provide high-quality genetic data for further investigations on IS.

In conclusion, the WGS and genome-wide analyses on CNSR-III would help to refine our understanding on the genetic contribution to IS/TIA and stroke outcomes, and possibly discover novel therapeutic targets for secondary prevention.

Author affiliations

- ¹Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China
- ²China National Clinical Research Center for Neurological Diseases, Beijing, China
- ³Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China
- ⁴Department of Neurology, The First people's Hospital of Taizhou, Taizhou, China
- ⁵Department of Neurology, Wuhan First Hospital, Wuhan, China
- ⁶Department of Neurology, The First People's Hospital of Xi'an, Xi'an, China
- ⁷Department of Neurology, The Second People's Hospital of Jinzhong, Jinzhong, China
- ⁸Department of Neurology, WuYuan County People's Hospital, Bayannur, China
- ⁹Department of Neurology, Qiqihar City Rongjian Stroke Prevention and Treatment Institute, Qiqihar, China
- ¹⁰Department of Neurology, Peking University Shenzhen Hospital, Shenzhen, China
- ¹¹Department of Neurology, Luoyang Central Hospital, Luoyang, China
- ¹²Department of Neurology, Qinghai Provincial People's Hospital, Xining, China
- ¹³Department of Neurology, Tianjin Huanhu Hospital, Tianjin, China
- ¹⁴Department of Neurology, Pingdingshan First People's Hospital, Pingdingshan, China
- ¹⁵Department of Neurology, Changzhi People's Hospital, Changzhi, China
- ¹⁶Department of Neurology, Zhejiang Provincial People's Hospital, Hangzhou, China
- ¹⁷Department of Neurology, Xinjiang Uyghur Autonomous Region People's Hospital, Urumqi, China

Twitter Yilong Wang @yilong

Contributors Study concept and design: SC, HaL and YoW. Drafting of the manuscript: SC, ZX and YL. Statistical analysis: AW, XH and ZX. Study supervision and organisation of the project: JL, YJ, XM, HaL, YIW and YoW. Supplying patients: ZW, GC, SW, ZJ, YC, XQ, JW, BS, WJ, ZA, WX, LZ, YG and HoL.

Funding This study was supported by grants from the Ministry of Science and Technology of the People's Republic of China (2016YFC0901002, 2016YFC0901001), Beijing Municipal Science & Technology Commission (D171100003017002), Beijing Municipal Administration of Hospitals' Mission Plan (SML20150502) and National Science and Technology Major Project (2017ZX09304018). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study was approved by the ethics committees of Beijing Tiantan Hospital and all other research centres according to the principles expressed in the Declaration of Helsinki.

Provenance and peer review Not commissioned; internally peer reviewed.

Data availability statement Data are available upon reasonable request. Data in this article are available upon reasonable request.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Anxin Wang <http://orcid.org/0000-0003-4351-2877>

Hao Li <http://orcid.org/0000-0002-8591-4105>

Yongjun Wang <http://orcid.org/0000-0002-9976-2341>

REFERENCES

- 1 GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2018;392:1736-88.
- 2 Wang Y, Li Z, Wang Y, et al. Chinese stroke center alliance: a national effort to improve healthcare quality for acute stroke and transient ischaemic attack: rationale, design and preliminary findings. *Stroke Vasc Neurol* 2018;3:256-62.
- 3 Bersano A, Markus HS, Quagliani S, et al. Clinical Pre-genetic screening for stroke monogenic diseases: results from Lombardia GENS registry. *Stroke* 2016;47:1702-9.
- 4 Malik R, Chauhan G, Traylor M, et al. Multi-ancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* 2018;50:524-37.
- 5 NINDS Stroke Genetics Network (SiGN), International Stroke Genetics Consortium (ISGC)Pulit SL, McArdle PF, Wong Q. Loci associated with ischaemic stroke and its subtypes (sign): a genome-wide association study. *Lancet Neurol* 2016;15:174-84.
- 6 Neurology Working Group of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, the Stroke Genetics Network (SiGN), and the International Stroke Genetics Consortium (ISGC). Identification of additional risk loci for stroke and small vessel disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 2016;15:695-707.
- 7 Traylor M, Farrall M, Holliday EG, et al. Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol* 2012;11:951-62.
- 8 Holliday EG, Maguire JM, Evans T-J, et al. Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. *Nat Genet* 2012;44:1147-51.
- 9 International Stroke Genetics Consortium (ISGC), Wellcome Trust Case Control Consortium 2 (WTCCC2), Bellenguez C, et al. Genome-Wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet* 2012;44:328-33.
- 10 Dichgans M, Pulit SL, Rosand J. Stroke genetics: discovery, biology, and clinical applications. *Lancet Neurol* 2019;18:587-99.
- 11 Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell* 2019;177:26-31.
- 12 Wang Y, Jing J, Meng X, et al. The third China national stroke registry (CNSR-III) for patients with acute ischaemic stroke or transient ischaemic attack: design, rationale and baseline patient characteristics. *Stroke Vasc Neurol* 2019;4:158-64.
- 13 Auton A, Brooks LD, 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. *Nature* 2015;526:68-74.
- 14 Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884-90.
- 15 Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;43:11.10.1-11.10.33.
- 16 Eggertsson HP, Kristmundsdottir S, Beyter D, et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun* 2019;10:5402.
- 17 Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;32:1220-2.
- 18 Chang CC, Chow CC, Tellier LC, et al. Second-Generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.
- 19 Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;26:2867-73.
- 20 Hubisz MJ, Falush D, Stephens M, et al. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* 2009;9:1322-32.
- 21 McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol* 2016;17:122.
- 22 Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am J Hum Genet* 2017;100:267-80.
- 23 GTEx Consortium. Human genomics. The Genotype-Tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;348:648-60.
- 24 Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012;40:D930-4.

- 25 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, *et al.* Predicting splicing from primary sequence with deep learning. *Cell* 2019;176:535–48.
- 26 Lappalainen I, Lopez J, Skipper L, *et al.* DbVar and DGVa: public Archives for genomic structural variation. *Nucleic Acids Res* 2013;41:D936–41.
- 27 MacDonald JR, Ziman R, Yuen RKC, *et al.* The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;42:D986–92.
- 28 Kishikawa T, Momozawa Y, Ozeki T, *et al.* Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci Rep* 2019;9:1784.
- 29 Rashkin S, Jun G, Chen S, *et al.* Optimal sequencing strategies for identifying disease-associated singletons. *PLoS Genet* 2017;13:e1006811.
- 30 Bentley DR, Balasubramanian S, Swerdlow HP, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53–9.