www.nature.com/scientificdata

SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

Reference exome data for a Northern Brazilian population

Alexia L. Weeks[1,5], Richard W. Francis[1,5], Joao I. C. F. Neri[2], Nathaly M. C. Costa[2], Nivea M. R. Arrais[3], Timo Lassmann[1,6], Jenefer M. Blackwell[1,6] ✉ & Selma M. B. Jeronimo[2,4,6]

Exome sequencing is widely used in the diagnosis of rare genetic diseases and provides useful variant data for analysis of complex diseases. There is not always adequate population-specific reference data to assist in assigning a diagnostic variant to a specific clinical condition. Here we provide a catalogue of variants called after sequencing the exomes of 45 babies from Rio Grande do Nord in Brazil. Sequence data were processed using an 'intersect-then-combine' (ITC) approach, using GATK and SAMtools to call variants. A total of 612,761 variants were identified in at least one individual in this Brazilian Cohort, including 559,448 single nucleotide variants (SNVs) and 53,313 insertion/deletions. Of these, 58,111 overlapped with nonsynonymous (nsSNVs) or splice site (ssSNVs) SNVs in dbNSFP. As an aid to clinical diagnosis of rare diseases, we used the American College of Medicine Genetics and Genomics (ACMG) guidelines to assign pathogenic/likely pathogenic status to 185 (0.32%) of the 58,111 nsSNVs and ssSNVs. Our data set provides a useful reference point for diagnosis of rare diseases in Brazil. (169 words).

Background & Summary

Next-generation sequencing of protein-coding regions, known as whole exome sequencing (WES), has enabled molecular diagnoses for thousands of rare disease patients (reviewed[1]) and provides useful variant data for genetic studies of complex diseases. As the use of this technology spreads world-wide it is becoming more important to understand genetic heterogeneity at a population-specific level, and to generate adequate population-specific reference data to assist clinical geneticists in assigning a diagnostic variant to a specific clinical condition. One region in which this is becoming increasingly important is in South America, and more specifically in Brazil, where genetic causes of clinical traits such as congenital microcephaly, ocular disease, need to be differentially diagnosed from those associated with Zika virus infection acquired *in utero*. In examining a cohort of 45 Brazilian babies from the State of Rio Grande do Norte we undertook WES to ascertain that none of 44 babies presenting with Zika-associated microcephaly were due to pathogenic genetic variants known to be associated with this clinical trait. One baby presented with familial congenital microcephaly. Here we describe the baseline data on variants identified in this population, with a specific focus on known and novel (i.e. those exclusive to this Brazilian population) rare variants that will inform the diagnosis of rare genetic diseases in Brazil. The data also provides useful information on novel and common variants that add to our knowledge of genetic heterogeneity in Brazil and may contribute to studies of genetic risk factors for complex diseases.

The exome data were processed with GATK 4.0.2.0[2,3] and SAMtools 1.7[4] using an 'intersect-then-combine' (ITC) approach. Variant calling was performed using the GATK best practices and SAMtools mpileup, only variants identified by both methods were retained. We calculated an average sequence depth of 97.4% at 20X coverage and 94.4% at 30X coverage (Fig. 1a). An average transition/transversion (Ts/Tv) ratio of 2.30 was observed for the Brazilian sample used here (Fig. 1b).

Sequences were aligned to the hg19 reference human genome and a total number of 612,761 variants were identified in at least one individual. Of these variants, 559,448 were single nucleotide variants (SNVs) and 53,313

[1]Telethon Kids Institute, The University of Western Australia, Perth Children's Hospital, Western Australia, Perth, Australia. [2]Institute of Tropical Medicine of Rio Grande do Norte and Department of Biochemistry, Universidade Federal do Rio Grande do Norte, Natal, Rio de Grande do Norte, Natal, Brazil. [3]Department of Pediatrics, Federal University of Rio Grande do Norte and Empresa Brasileira de Serviços Hospitalares, Natal, Brazil. [4]National Institute of Science and Technology of Tropical Diseases, Natal, RN, Brazil. [5]These authors contributed equally: Alexia L. Weeks, Richard W. Francis. [6]These authors jointly supervised this work: Timo Lassmann, Jenefer M. Blackwell, Selma M. B. Jeronimo. ✉e-mail: jenefer.blackwell@telethonkids.org.au

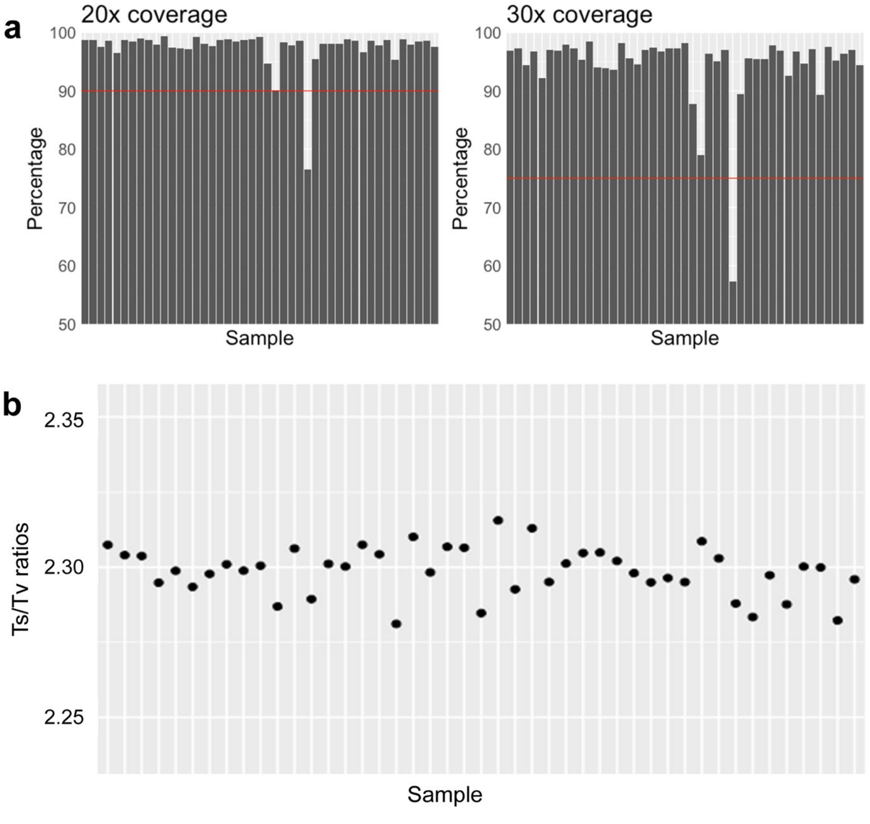

Fig. 1 (**a**) WES coverage for the at 20X and 30X depth. Each bar represents an individual sample and the percentage of bases with at least 20X or 30X coverage. The red lines mark the 90% and 75% coverages at 20X and 30X depths, respectively, which are optimal targets for WES that most of the samples achieved. (**b**) Ts/Tv ratio calculated individually for all individuals using SNVs passing the GATK best practice VQSR threshold.

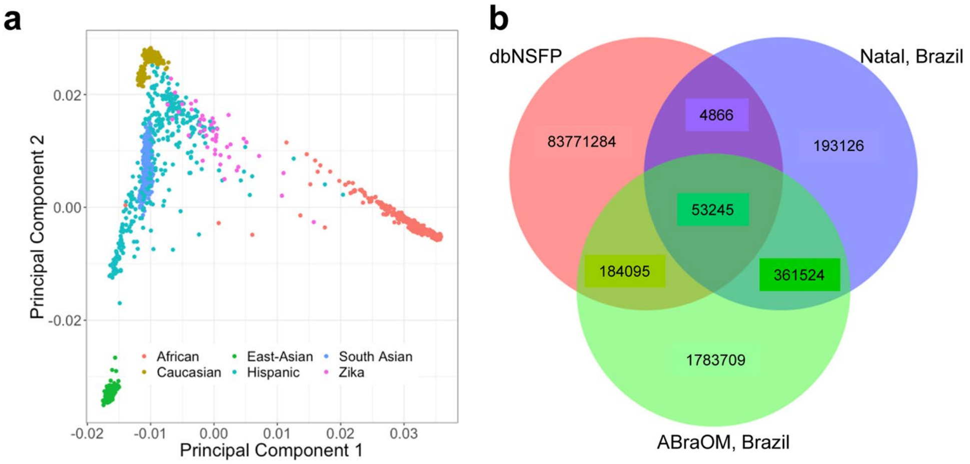

Fig. 2 (**a**) PCA plot to demonstrate ethnic admixture in the Brazilian sample compared to 1000 G populations grouped as African (ACB, ASW, ESN, GWD, LWK, MSL, YRI); Hispanic (CLM, MXL, PEL, PUR); East-Asian (CDX, CHB, CHS, JPT, KHV); South Asian (BEB, GIH, ITU, PJL, STU); and Caucasian (CEU, FIN, GBR, IBS, TSI). (**b**) Venn diagram showing overlap between SNVs in two Brazilian samples (ABraOM database of exome variants from 609 elderly Brazilians from Sao Paulo State[5] and the 45 exomes from Rio de Grande do Norte State studied here) compared to large public domain databases.

were insertions/deletions (indels). To evaluate admixture in this Brazilian sample we carried out principal component analysis on an LD-pruned set of SNVs with minor allele frequencies >0.1. Comparison with 1000 G populations indicated predominant admixture between Caucasian and Negroid populations (Fig. 2a), consistent with data from the ABraOM database of exome variants from 609 elderly Brazilians from Sao Paulo State[5]. In

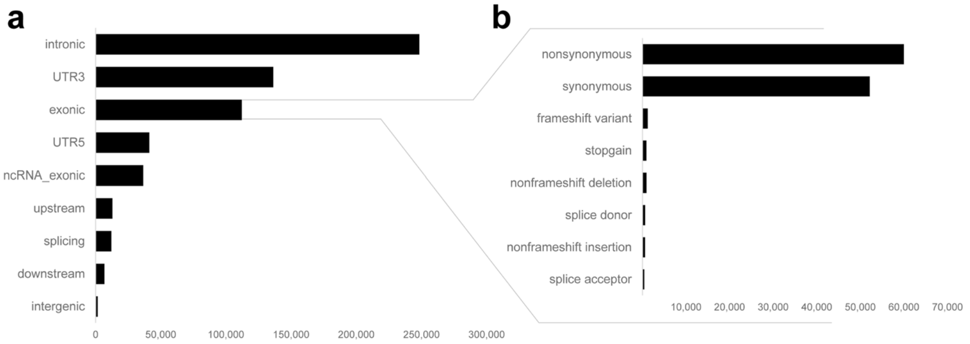

Fig. 3 Annotation of identified variants reported by (**a**) genomic location and (**b**) the main variant consequences. Other variant consequences not in the figure included: stop_lost (164); start_lost (219); protein_altering_variant (4); incomplete_terminal_codon_variant (8); stop_retained_variant (72); coding_sequence_variant (16); and mature_miRNA_variant (160).

comparing our data with the ABraOM database we found 414,769 variants in common with the ABraOM study and 197,992 that were unique to our study sample. Comparing the data with large public domain datasets (dbSNP 151[6], 1000 Genomes Phase 3[7], TWINSUK[8], ESP6500[9], UK10K[10], ExAC[11] and gnomAD[12] databases) we found 361,524 variants that were unique to the combined Brazilian datasets (Fig. 2b).

The 612,761 variants in our study sample were annotated with VEP[13] to provide variant types and consequences (Fig. 3). Most variants were categorised as intronic, exonic or UTR3, consistent with design of the exome sequencing capture kit. A total of 248,329 intronic variants, 117,524 exonic variants and 136,266 UTR3 variants were present.

Exome variants of interest in diagnosis of rare genetic diseases usually fall within the categories of nonsynonymous SNVs (nsSNVs) and splice-site variants (ssSNVs). To identify this potentially functional subset of variants in our dataset we looked for overlap between our variants and those present in the dbNSFP v4.0[14,15] database of human nsSNVs and ssSNVs. Using the search_dbNSFP40a function we identified 58,111 nsSNVs/ssSNVs in our sample that were present in dbNSFP.

To further identify nsSNVs/ssSNVs that may be pathogenic for genetic diseases we determined the number that classify as "pathogenic" or "likely pathogenic" according to the American College of Medicine Genetics and Genomics (ACMG) standards and guidelines[16] (Supplementary Table 1). Of the 58,111 nsSNVs/ssSNVs in our sample a total of 12 (0.02%) were classified as "pathogenic" and 173 (0.30%) as "likely pathogenic". Details of these variants is provided in Supplementary Table 2.

Methods

Study population. Subjects were recruited through the Pediatric Hospital of the Federal University of Rio Grande do Norte or through visits to households that had cases suspected of microcephaly in Natal and other cities where cases of microcephaly were reported during the 2015–2016 ZIKV outbreak, Rio Grande do Norte, Brazil. The sample comprised 45 babies (26 males aged mean ± SD 25.50 ± 7.17 months; 19 females aged mean ± SD 24.79 ± 6.73 months), 44 with confirmed Zika-associated congenital microcephaly and one baby with familial congenital microcephaly. None of the babies with Zika-related microcephaly had a deleterious genetic variant previously known to be associated with genetically determined congenital microcephaly that could account for their phenotype (see Supplementary Table 2). Nor did we find a variant that matched deleterious variants in dbNSFP v4.0 that would account for the one familial case of microcephaly. The complete list of genes and filtering strategy that we applied to look for microcephaly variants is provided in Supplementary Table 3.

Ethical considerations. This study was undertaken with ethical approval from the institutional review board of the Universidade Federal do Rio Grande do Norte/Comissão Nacional de Ética em Pesquisa (CAAE 53111416.7.0000.5537). Written consent was obtained from the parents or legal guardians of babies who ranged in age from 5 months to 40 months. The individual consent included an option to accept or refuse continued use of their genetic or clinical data in further studies. The parents or legal guardian of all subjects included in the study had given consent for storage and future use of deidentified DNA samples and data for their children.

Whole exome sequencing. The DNA samples were prepared following the Agilent SureSelect XT + UTR v6 protocol and sequenced on a HiSeq. 4000 system using 150 bp paired end chemistry at the Genomics Division, Iowa Institute of Human Genetics, University of Iowa, USA. Sequence data was processed with GATK 4.0.2.0[2,3] and SAMtools 1.7[4] using an 'intersect-then-combine' (ITC) approach. Variant calling was performed with GATK following best practices[17] and with SAMtools[4] using the mpileup function. Only variants identified by both methods were retained. Sequence coverage was calculated using BEDtools[18] with the -d parameter to calculate the per base depth and then the percentage of bases with at least 20X and 30X coverage were calculated.

Variant annotation. Prior to annotation, the data were normalized and decomposed with VT v0.57721[19]. Variant annotation was performed using the Variant Effect Predictor (VEP v97.3)[13]. VEP annotated variants as splicing, ncRNA, UTR5, UTR3, intronic, upstream, downstream or intergenic, with exonic variants categorised as start lost, stop lost, stop gain, frameshift insertion, frameshift deletion, nonframeshift insertion, nonframeshift deletion, protein altering variant, incomplete terminal codon variant, stop retained variant, coding sequence variant, mature miRNA variant, missense SNV or synonymous SNV.

Overlap with known variants. Variants in our dataset were compared with data from the Brazilian ABraOM databases[5] and with large public domain datasets dbSNP 151[6], 1000 Genomes Phase 3[7], TWINSUK[8], ESP6500[9], UK10K[10], ExAC[11] and gnomAD[12] databases. To gain a handle on functionality relevant to rare disease diagnosis, variants in our sample were also compared to the 84,013,490 nsSNVs and ssSNVs present in the dbNSFP[14,15] v4.0 database.

Principal component analysis. Scripts used to perform the data processing and plotting for principal component analysis are available at https://github.com/richardwfrancis/zika_admixture.

Classification of variant pathogenicity. To further identify nsSNVs/ssSNVs that may be pathogenic for genetic diseases we determined whether they classified as "pathogenic" or "likely pathogenic" according to the ACMG standards and guidelines[16] using criteria laid out in Supplementary Table 1 and implemented using the python script available here: https://github.com/TimoLassmann/Phenoparser/blob/master/scripts/acmg.py.

Data Records

The full set of variants has been recorded as two VCF files: (i) The complete normalised and VEP annotated version containing only PASSed variants and (ii) the subset of variants found in dbNSFP, which is further annotated with ACMG classification. These files have been deposited in the European GenomePhenome Archive (EGA) under the accession number EGAS00001004112[20]. Summary allele frequency data have also been deposited in the European Nucleotide Archive (ENA) under the project and analysis accession numbers PRJEB39409 and ERZ1466912, respectively[21].

Technical Validation

The Ts/Tv ratio was calculated for each sample using the stats function in BCFtools as a quality control metric. It has been reported that a Ts/Tv ratio of 2.8 is expected for WES[22]. However this varies greatly by genome region and functionality[23], and is usually ~3.0 for exome regions and ~2.0 outside of exome regions[24].

Usage Notes

Summary allele frequency data are available as open access through the ENA. The variant data are made available through the EGA. Access to variant data will be granted to qualified researchers for appropriate health related uses, subject to review by a study-specific Data Access Committee (DAC).

Received: 23 January 2020; Accepted: 14 August 2020;
Published online: 21 October 2020

References

1. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* **19**, 253–268 (2018).
2. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–303 (2010).
3. Poplin, R. *et al*. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2018).
4. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
5. Naslavsky, M. S. *et al*. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum Mutat* **38**, 751–763 (2017).
6. Sherry, S. T. *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–11 (2001).
7. The Genomes Project, C. *et al*. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
8. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin research and human genetics: the official journal of the International Society for Twin Studies* **16**, 144–149 (2013).
9. Exome Variant Server. NHLBI GO exome sequencing project (ESP), Seattle, WA (http://evs.gs.washington.edu/EVS/ (accessed 7 June 2020).
10. Consortium, U. K. *et al*. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
11. Karczewski, K. J. *et al*. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research* **45**, D840–D845 (2017).
12. Lek, M. *et al*. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (2016).
13. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
14. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894–9 (2011).
15. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235–41 (2016).
16. Richards, S. *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–24 (2015).
17. Van der Auwera, G. A. *et al*. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1–33 (2013).
18. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
19. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
20. Weeks, A., Francis, R.W., Blackwell, J.M. & Jeronimo, S.M.B. Reference exome data for a Northern Brazilian population. *European Genome-phenome Archive* http://identifiers.org/ega.study/EGAS00001004112 (2020).
21. Weeks, A., Francis, R.W., Blackwell, J.M. & Jeronimo, S.M.B. Reference exome data for a Northern Brazilian population. *European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB39409 (2020).
22. Carson, A. R. *et al*. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* **15**, 125–125 (2014).

23. Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* **31**, 318–323 (2015).
24. Bainbridge, M. N. *et al*. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biology* **12**, R68–R68 (2011).

Acknowledgements

The work was partially supported by a Wesfarmers Seed Grant held by JB at the Telethon Kids Institute in Western Australia and by CAPES (Grant number 440893/2016-0) and CNPq (Grant number 88881.130729/2016-01) in Brazil.

Author contributions

A.L.W. analysed and interpreted the exome data sets and assisted in drafting the manuscript. R.W.F. assisted with variant calling, prepared PCA plots, undertook the analysis of ACMG pathogenicity of variants, and assisted in drafting the manuscript. J.I.C.F.N. conducted field work, subject recruitment and clinical ascertainment. F.P.F.N. collected and processed blood and performed quality control of samples. N.M.C.C. data entering. N.M.R.A. aided patient recruitment and ascertainment of phenotypes. T.L. supervised the exome analysis. J.M.B. co-led the study and drafted the manuscript. S.M.B. was lead investigator of the study in Brazil, assisted with clinical evaluation of participants and sample collection, and assisted with preparation of the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41597-020-00703-y.

Correspondence and requests for materials should be addressed to J.M.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.