




SCIENTIFIC DATA



OPEN

DATA DESCRIPTOR

Reference exome data for Australian Aboriginal populations to support health-based research

Alexia L. Weeks¹, Heather A. D'Antoine², Melita McKinnon², Genevieve Syn¹, Dawn Bessarab³, Ngiare Brown⁴, Steven Y. C. Tong ^{2,5}, Bo Reményi², Andrew Steer⁶, Lesley-Ann Gray^{7,8}, Michael Inouye^{7,8}, Jonathan R. Carapetis¹, Jenefer M. Blackwell ^{1,9} & Timo Lassmann ^{1,9} ✉

Whole exome sequencing (WES) is a popular and successful technology which is widely used in both research and clinical settings. However, there is a paucity of reference data for Aboriginal Australians to underpin the translation of health-based genomic research. Here we provide a catalogue of variants called after sequencing the exomes of 50 Aboriginal individuals from the Northern Territory (NT) of Australia and compare these to 72 previously published exomes from a Western Australian (WA) population of Martu origin. Sequence data for both NT and WA samples were processed using an 'intersect-then-combine' (ITC) approach, using GATK and SAMtools to call variants. A total of 289,829 variants were identified in at least one individual in the NT cohort and 248,374 variants in at least one individual in the WA cohort. Of these, 166,719 variants were present in both cohorts, whilst 123,110 variants were private to the NT cohort and 81,655 were private to the WA cohort. Our data set provides a useful reference point for genomic studies on Aboriginal Australians.

Background & Summary

Whole exome sequencing (WES) is a popular and successful technology which is becoming more widely used in both research and clinical settings^{1,2}. This technology is a fraction of the cost of other methods such as whole genome sequencing (WGS) and provides informative data on common variants found widely in the general population, as well as novel variants and variants involved in genetic diseases¹.

As there are currently limited data available for Aboriginal Australians to be used as a reference for health-based research, a study in partnership with two Aboriginal Australian populations was carried out. The first population living at the edge of the Western Desert in Western Australia (WA) includes 72 Aboriginal Australians and has been previously described³. The second population includes 50 Aboriginal Australians from the Northern Territory (NT) of Australia selected as healthy individuals representative of populations studied in a recent genome-wide association study of rheumatic heart disease (RHD)⁴. WES of these individuals was carried out to provide a reference data set for known and novel variants (i.e. those exclusive to these Aboriginal Australian populations). In parallel, we updated our variant calling pipeline in line with current developments in the field⁵, and have re-called variants from the WA population to provide an improved and more concise set of variants. Together these two population data sets expand our Aboriginal Australian reference panel for use in the health sector, with particular reference to rare variants that will inform the diagnosis of rare genetic diseases in

¹Telethon Kids Institute, The University of Western Australia, Perth Children's Hospital, Perth, Western Australia, Australia. ²Menzies School of Health Research, Charles Darwin University, Darwin, Northern Territory, Australia.

³Centre for Aboriginal Medical and Dental Health, The University of Western Australia, Crawley, Western, Australia.

⁴School of Education, The University of Wollongong, New South Wales, Australia. ⁵Victorian Infectious Disease Service, The Royal Melbourne Hospital, and Doherty Department, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Victoria, Australia. ⁶Group A Streptococcal Research Group, Murdoch

Childrens Research Institute, Melbourne, Victoria, Australia and Centre for International Child Health, Department

of Paediatrics, Royal Children's Hospital, Melbourne, Victoria, Australia. ⁷Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia. ⁸Department of Public Health and Primary Care, The University of Cambridge,

Cambridge, UK. ⁹These authors contributed equally: Jenefer Blackwell, Timo Lassmann. ✉e-mail: timo.lassmann@telethonkids.org.au

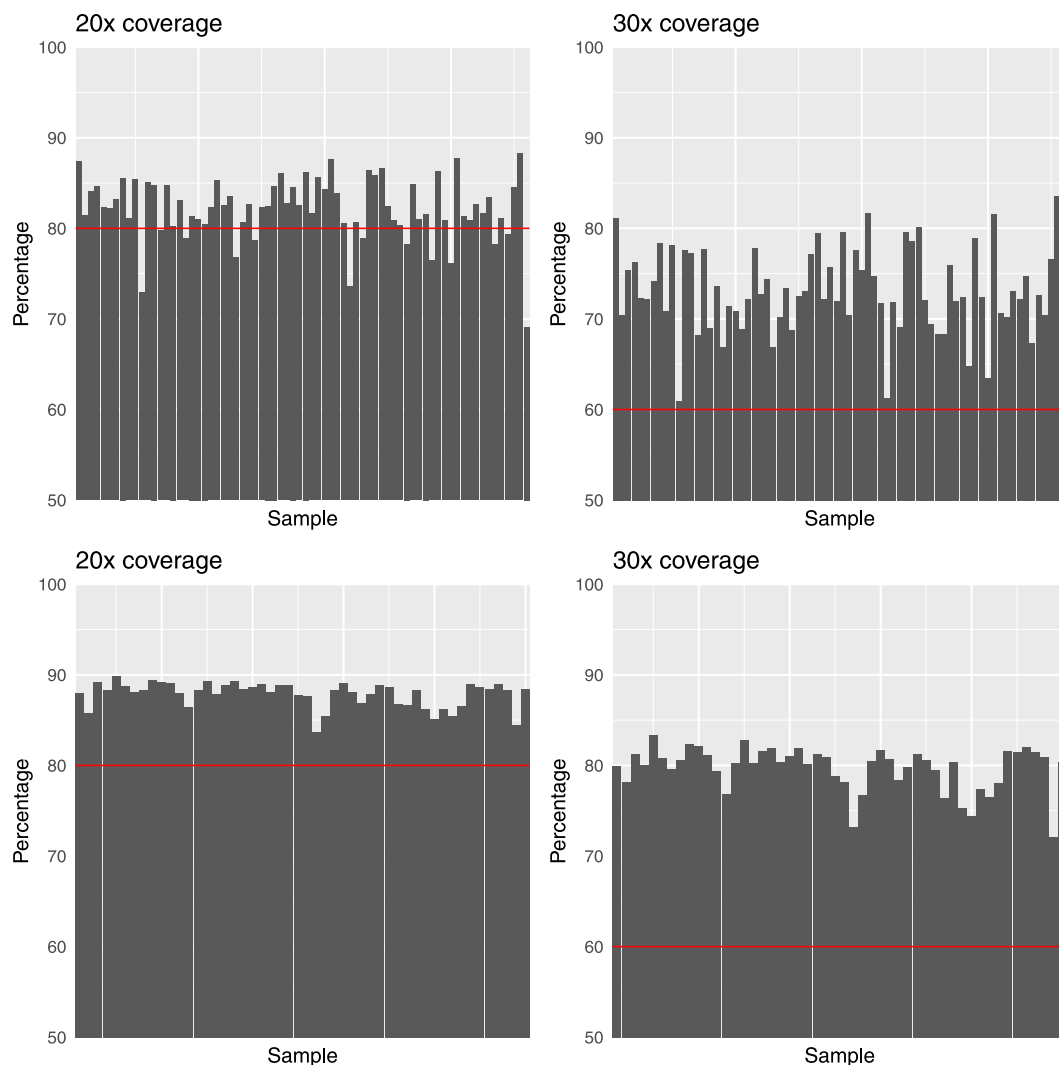


Fig. 1 Whole exon coverage statistics. Panels show coverage at 20X and 30X for the WA population (**a,b**) and 20X and 30X coverage for the NT population (**c,d**). Each bar represents an individual sample and the percentage of bases with at least 20X or 30X coverage. Red lines mark 80% and 60% coverage at 20X and 30X depths, respectively, which all NT samples and most WA samples achieve.

Aboriginal Australians⁶. In addition, our original dataset³ has been used to undertake the first systematic assessment of blood group antigen profiles in Indigenous Australian which is expected to guide transfusion practice for Aboriginal Australians⁷. For these health-based applications our data have some advantages over published⁸ WGS data for Aboriginal Australians where small numbers of DNAs (n of 5 to 13) from geographically and ethnically disparate groups across Australia were employed.

The exome data were processed with GATK 4.0.2.0^{9,10} and SAMtools 1.7¹¹ using an ‘intersect-then-combine’ (ITC) approach, where variant calling was performed with GATK following the best practices, and also performed with SAMtools using the mpileup function, and only variants identified by both methods were retained. The NT population had an average sequence depth of 89.8% at 20X coverage and 83.3% at 30X coverage. The WA population had an average sequence depth of 82.2% at 20X coverage and 72.9% at 30X coverage (Fig. 1).

The GATK pipeline involved several quality recalibration steps to ensure the quality of the variants called. The base quality score recalibration (BQSR) step is designed to detect systematic errors by the sequencer when it is calling base quality in the pre-processing stage. Later, the variant quality score recalibration (VQSR) calculates a new quality score named the VQSLOD (for variant quality score log-odds), which allows variant filtering with a balance of sensitivity and specificity to identify real variants whilst limiting false variants.

Sequences were aligned to the hg19 reference human genome and a total number of 289,829 variants were identified in at least one individual in the NT cohort and 248,374 variants in at least one individual in the WA cohort. Of these, 166,719 variants were present in both cohorts, whilst 123,110 variants were private to the NT cohort and 81,655 were private to the WA cohort (Fig. 2).

The variants identified in these cohorts were compared to single nucleotide variants (SNVs) recorded in the database for nonsynonymous SNVs’ functional predictions (dbNSFP v3.5)^{2,12}. This dbNSFP is a database of

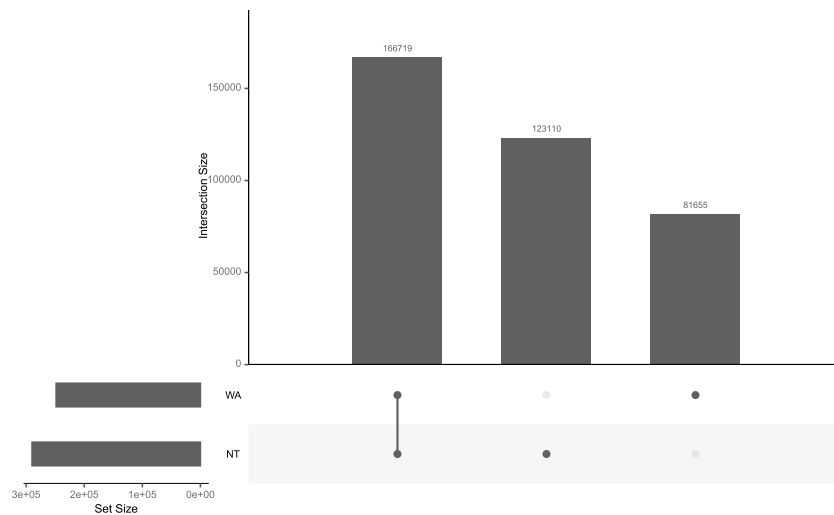


Fig. 2 Matrix layout for the intersections of variant identified in the WA and NT populations relative to the hg19 human reference genome. Dark circles in the matrix indicate sets that are part of the intersection.



Fig. 3 Annotation of identified variants (a) and consequence (b) for the WA and NT populations.

human SNPs that includes 20 functional prediction scores as well as 6 conservation scores and allele frequency data from 5 datasets for the 82,832,027 nonsynonymous SNVs (nsSNVs) and splice-site variants (ssSNVs) in the human genome. We identified 8,801 and 12,363 nsSNVs not present in dbNSFP in the WA and NT population respectively. Of these variants, 2,356 were present in both populations.

Variants were annotated with ANNOVAR¹³ to provide gene names and variant types and consequences (Fig. 3). The majority of variants were located in introns and exons for both populations. Totals of 83,585 and 90,642 exonic variants were present for the WA and NT populations, respectively. Whilst exons are the target regions in exome sequencing, a number of variants are also captured from outside of these target regions. A target padding of 100 bp was applied to the targets as recommended in the best practices. This target padding allows the inclusion of flanking regions so variants just outside of the target regions can be called. Exonic variants were further characterised by their functional consequence. The majority of these variants were nonsynonymous and synonymous SNVs for both populations.

Methods

Study populations. Subjects were recruited from Aboriginal Australian communities in the NT and in WA as described previously^{3,4}. The NT individuals were selected randomly and included samples from 16 of the 19 communities originally studied⁴. All had given consent for storage and future use of deidentified DNA samples and data. The 50 individuals comprised 29 controls (11 males aged 20–58 years; 18 females aged 18–64 years) and 21 RHD cases (7 males aged 20–52 years; 14 females aged 20–60 years). These were a subset of individuals used in a genome-wide study of genetic risk factors for RHD⁴. The WA sample comprised 72 individuals from an Aboriginal Australian community of Martu ancestry for whom exome sequence data were already available as previously described³. This was a subset of individuals used in a genome-wide study of genetic risk factors for body mass index and type 2 diabetes¹⁴.

Whole exome sequencing. The 50 NT DNA samples were prepared following the Agilent SureSelect XT Human All Exon + UTR v5 protocol and sequenced on an Illumina HiSeq. 2500 system outsourced to the

Australian Genome Resource Facility (AGRF). Exome sequence data from the 72 WA samples had been similarly obtained from DNA prepared following the Illumina TruSeq protocol³. Sequence data for both NT and WA samples were processed using an ‘intersect-then-combine’ (ITC) approach, using GATK⁹ according to the GATK Best Practices recommendations^{15,16} and SAMtools¹¹ using the mpileup function to call variants. Briefly, sequences were aligned to the hg19 reference genome with BWA-MEM¹⁷, followed by the removal of PCR duplicates and base quality score recalibration. Variant calling was performed using both callers and a single VCF file of the intersect was produced.

The coverage was calculated using BEDtools¹⁸ with the `-d` parameter to calculate the per 4base depth and then the percentage of bases with at least 20X and 30X coverage were calculated.

Overlapping with known variants. The VCF files from both populations were intersected using BCFtools¹¹; specifically using the `isec` command. This produced individual VCF files for variants unique to each cohort, as well as VCF files for the intersect of both cohorts.

Non-synonymous SNVs (nsSNVs) were compared to the 83,422,341 nsSNVs and splicing site SNVs (ssSNVs) present in the dbNSFP database. We excluded variants in the following publicly available databases to identify variants unique to the Aboriginal Australian populations: dbSNP 150¹⁹, 1000 Genomes Phase 3²⁰, TWINSUK²¹, ESP6500²², ExAC²³ and gnomAD²⁴.

Variant annotation. Variant annotation was performed using ANNOVAR¹³ (version 2018Apr16) to annotate variants. Specifically, the `table_annovar.pl` program was used to annotate variants against RefSeq sequences (version 20170601). Variants were annotated as exonic, splicing, ncRNA, UTR5, UTR3, intronic, upstream, downstream or intergenic. Exonic variants are further categorised as stoploss, stopgain, frameshift insertion, frameshift deletion, nonframeshift insertion, nonframeshift deletion, nonsynonymous SNV and synonymous SNV or unknown.

Data Records

The full set of variants for each population has been recorded as a single multi-sample VCF file. These files have been deposited in the EGA under the accession number EGAS00001003745²⁵.

Technical Validation

The transition/transversion (Ts/Tv) ratio was calculated for each sample using BCFtools, specifically the `stats` function, as a quality control metric. Transitions are interchanges between purines (A, G) or pyrimidines (C, T), and transversions are changes between purines and pyrimidines. These changes occur at a ratio of 0.5 when there is no bias towards either purines or pyrimidines. In human DNA, transitions are more frequently observed due to molecular mechanisms, such as tautomeric shifts²⁶, that lead to a bias in transversions of purines versus pyrimidines and as such, a ratio greater than 0.5 is expected.

It has been reported that a Ts/Tv ratio of 2.8 is expected for WES²⁷, however this varies greatly by genome region and functionality²⁸. It has been reported to be around 3.0 for exome regions, and about 2.0 outside of exome regions²⁹. An average Ts/Tv ratio of 2.42 and 2.39 was observed for the WA and NT populations respectively (Fig. 4). When examining only the exonic variants, the Ts/Tv ratio was 3.01 and 3.02 for the WA and NT populations respectively, and 2.17 for non-exonic variants in both data sets, which is consistent with reported Ts/Tv ratios²⁹. The Ts/Tv ratios presented are below the expected value of 2.8, however this is due to the large number of non-exonic variants captured. Whilst the target regions are exonic, the target probes have been designed to capture regions slightly outside of the exons to capture non-coding regions such as the 5' and 3' UTRs, which may have functional implications. Variants in these UTRs are not as constrained as those in exons, and as a result, transversions are more common and can lower the Ts/Tv ratio.

Usage Notes

The NT study was undertaken with ethical approval from the Human Research Ethics Committee (HREC) of the Northern Territory Department of Health and Menzies School of Health Research (ID HREC-2010-1484) and the Central Australian HREC (ID HREC-2014-241). The study was overseen by a project steering committee and 3 subcommittees: Aboriginal governance, clinical, and scientific, as previously described⁴. The individual consent incorporated an ‘opt-in’ design where participants selected which components of the study they were comfortable to participate in, and they were able to withdraw from the study at any stage. This included an option to accept or refuse continued use of their genetic or clinical data in further studies. All participants studied here accepted continued use of their genetic data. The vcf file for deidentified post-quality control data has been lodged in the European Genome-phenome Archive (accession number EGAS00001003745).

Ethical approval for the WA study was obtained from the Western Australian Aboriginal Health Ethics Committee (Reference 227 12/12). This ethics committee is responsible for reviewing health and medical research undertaken in Western Australian Aboriginal communities. It is registered with the National Health and Medical Research Council’s (NHMRC’s) Australian Health Ethics Committee (AHEC) and operates in accordance with the NHMRC National Statement on Ethical Conduct in Human Research 2007. The vcf file for de-identified data for 72 exomes re-analysed here has been lodged in the European Genome-phenome Archive (accession number EGAS00001003745).

The data for both NT and WA studies are made available through the European Genome-phenome Archive (EGA), subject to review by a study-specific Data Access Committee (DAC). This DAC is chaired by a leading Clinical Geneticist from Genetic Services of Western Australia, with membership including the Head of Aboriginal Research at the Telethon Kids Institute (TKI), the Associate Director for Aboriginal Programs at the Menzies School of Health Research, the Head of the Chronic & Severe Diseases Research Focus Area at TKI and Clinical Lead and Co-Director for Diabetes and Obesity Services at the Perth Children’s Hospital, the Director

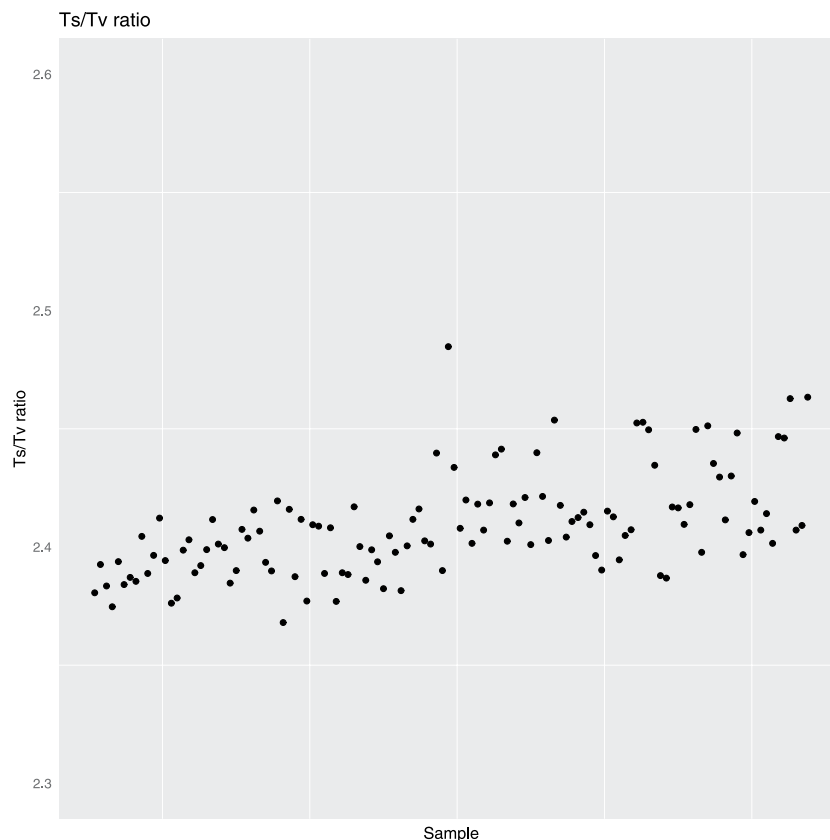


Fig. 4 Ts/Tv ratio calculated individually for all individuals using SNVs passing the VQSR threshold. The first 50 samples on the X-axis are the NT samples, the remainder are the WA samples. The differences in average Ts/Tv ratios between NT and WA samples reflects differences in the exonic/intronic sequence ratios in the two different capture panels employed for WES.

of the Office of Public Health Genomics for the WA Health Department, a leading Genetic Statistician from TKI, and a Senior Research Fellow in Aboriginal Health at TKI. The DAC is managed *ex officio* by a genetic ethicist. Access to data will be granted to qualified researchers for appropriate health related uses. A qualified researcher refers to a scientist, who is employed, or a student enrolled at, or legitimately affiliated with an academic, non-profit or government institution, or a commercial company performing Aboriginal health related diagnostic services. Applicants are asked to complete a basic application form (which includes a brief summary of the proposal, so that the DAC can determine if the planned usage falls within consents) and to agree to the terms and conditions laid out in the Data Access Agreement (DAA). The DAA must be signed by the applicant and the relevant Head of Department, Head of Institute, or equivalent. If applications include a named collaborator then the collaborator's Institution must sign and submit a separate Data Access Agreement. Review by the DAC takes 2–3 weeks and if the application is approved, access via the EGA is then arranged for the applicant (<https://www.ebi.ac.uk/ega/about/access>). The application form, data access agreement, and further information are available from our website: <https://www.telethonkids.org.au/aghs>.

Permission to lodge de-identified genotype and basic demographic data (broad geographical location, age, sex and phenotype information) in the EGA was obtained from the Board of the local Aboriginal Health Service in WA, and from the project steering committee in NT. It should be noted that these approvals are for use of the data in health-based research, and not for use in pure population genetics research. The data are provided as reference data for health-based research and translation in Aboriginal Australian communities.

Received: 25 September 2019; Accepted: 24 March 2020;

Published online: 29 April 2020

References

- Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research* **27**, 157–164, <https://doi.org/10.1101/gr.210500.116> (2017).
- 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. *Human mutation* **37**, 235–241, <https://doi.org/10.1002/humu.22932> (2016).
- Tang, D. *et al.* Reference genotype and exome data from an Australian Aboriginal population for health-based research. *Scientific data* **3**, 160023, <https://doi.org/10.1038/sdata.2016.23> (2016).
- Gray, L. A. *et al.* Genome-Wide Analysis of Genetic Risk Factors for Rheumatic Heart Disease in Aboriginal Australians Provides Support for Pathogenic Molecular Mimicry. *J.Infect.Dis.* **216**, 1460–1470, <https://doi.org/10.1093/infdis/jix497> (2017).

5. Callari, M. *et al.* Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Medicine* **9**, 35, <https://doi.org/10.1186/s13073-017-0425-1> (2017).
6. Baynam, G. *et al.* Indigenous Genetics and Rare Diseases: Harmony, Diversity and Equity. *Adv Exp Med Biol* **1031**, 511–520, https://doi.org/10.1007/978-3-319-67144-4_27 (2017).
7. Schoeman, E. M., Roulis, E. V., Perry, M. A., Flower, R. L. & Hyland, C. A. Comprehensive blood group antigen profile predictions for Western Desert Indigenous Australians from whole exome sequence data. *Transfusion* **59**, 768–778, <https://doi.org/10.1111/trf.15047> (2019).
8. Malaspinas, A. S. *et al.* A genomic history of Aboriginal Australia. *Nature* **538**, 207–214, <https://doi.org/10.1038/nature18299> (2016).
9. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303, <https://doi.org/10.1101/gr.107524.110> (2010).
10. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, <https://doi.org/10.1101/201178> (2018).
11. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).
12. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation* **32**, 894–899, <https://doi.org/10.1002/humu.21517> (2011).
13. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164–e164, <https://doi.org/10.1093/nar/gkq603> (2010).
14. Anderson, D. *et al.* First genome-wide association study in an Australian aboriginal population provides insights into genetic risk factors for body mass index and type 2 diabetes. *Plos One* **10**, e0119333, <https://doi.org/10.1371/journal.pone.0119333> (2015).
15. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics/editorial board, Andreas D. Baxeavanis... [et al.]* **43**, 11 10 11-33, <https://doi.org/10.1002/0471250953.bi1110s43> (2013).
16. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–498, <https://doi.org/10.1038/ng.806> (2011).
17. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324> (2009).
18. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842, <https://doi.org/10.1093/bioinformatics/btq033> (2010).
19. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
20. The Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (2015).
21. 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, <https://doi.org/10.1017/thg.2012.89> (2013).
22. Shields, E. D., Russell, D. A. & Pericak-Vance, M. A. Genetic epidemiology of the susceptibility to leprosy. **79**, 1139–1143 (1987).
23. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research* **45**, D840–D845, <https://doi.org/10.1093/nar/gkw971> (2017).
24. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285, <https://doi.org/10.1038/nature19057> (2016).
25. *European Genome-phenome Archive*, <https://identifiers.org/ega.dataset:EGAD00001005189> (2020).
26. Griffiths, A. J. F. *et al.* In *An Introduction to Genetic Analysis* (2000).
27. Carson, A. R. *et al.* Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* **15**, 125–125, <https://doi.org/10.1186/1471-2105-15-125> (2014).
28. 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 (Oxford, England)* **31**, 318–323, <https://doi.org/10.1093/bioinformatics/btu668> (2015).
29. 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, <https://doi.org/10.1186/gb-2011-12-7-r68> (2011).

Acknowledgements

We acknowledge all investigators of these studies, the project teams including the local Aboriginal Health services, community-based researchers, the communities, agencies, and all the participants for their invaluable contribution to this project. The work was funded by grants APP634301 and APP1023462 from the Australian National Health and Medical Research Council to J.M.B. and J.R.C., respectively.

Author contributions

A.L.W. analysed and interpreted the data sets and drafted the manuscript. H.A.D. and M.M. managed the NT project, including management of ethical, legal, and social aspects of the study. M.M. carried out the field work and sample collection. D.B. and N.B. made significant contributions to governance and helped design the community engagement arms of the NT project. S.Y.C.T., B.R. and A.S. provided the major clinical inputs for diagnosis and review of patient records in the NT. G.S. prepared the DNAs, including quality control, and liaised with providers for exome capture and sequence-analysis, for both WA and NT studies. L.-A.G. and M.I. managed the study data and carried out randomised selection of individuals for the NT exome study. J.R.C. was lead investigator on the NT project. J.M.B. was lead investigator on the WA study, co-led all genetic aspects of the NT study, and undertook revisions of the manuscript. T.L. supervised the exome analysis and undertook revisions of the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.L.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2020