BMC Genomic Data

DATA NOTE

Open Access

DanMAC5: a browser of aggregated sequence variants from 8,671 whole genome sequenced Danish individuals

Karina Banasik[1*], Peter L. Møller[2], Tanya R. Techlo[3], Peter C. Holm[1], G. Bragi Walters[4], Andrés Ingason[5], Anders Rosengren[5], Palle D. Rohde[6], Lisette J. A. Kogelman[3], David Westergaard[1], Troels Siggaard[1], Piotr J. Chmura[1], Mona A. Chalmer[3], Ólafur Þ. Magnússon[4], Guðmundur Á. Þórisson[4], Hreinn Stefánsson[4], Daníel F. Guðbjartsson[4], Kári Stefánsson[4], Jes Olesen[3], Simon Winther[7], Morten Bøttcher[7], Søren Brunak[1], Thomas Werge[5†], Mette Nyegaard[2,6†] and Thomas F. Hansen[1,3†]

Abstract

Objectives Allele counts of sequence variants obtained by whole genome sequencing (WGS) often play a central role in interpreting the results of genetic and genomic research. However, such variant counts are not readily available for individuals in the Danish population. Here, we present a dataset with allele counts for sequence variants (single nucleotide variants (SNVs) and indels) identified from WGS of 8,671 (5,418 females) individuals from the Danish population. The data resource is based on WGS data from three independent research projects aimed at assessing genetic risk factors for cardiovascular, psychiatric, and headache disorders. To enable the sharing of information on sequence variation in Danish individuals, we created summarized statistics on allele counts from anonymized data and made them available through the European Genome-phenome Archive (EGA, https://identifiers.org/ega.dataset:EGAD00001009756) and in a dedicated browser, DanMAC5 (available at www.danmac5.dk). The summary level data and the DanMAC5 browser provide insight into the allelic spectrum of sequence variants segregating in the Danish population, which is important in variant interpretation.

Data description Three WGS datasets with an average coverage of 30x were processed independently using the same quality control pipeline. Subsequently, we summarized, filtered, and merged allele counts to create a high-quality summary level dataset of sequence variants.

Keywords Whole genome sequencing, Variant interpretation, Sequence variants, Minor allele counts, Browser

†Thomas Werge, Mette Nyegaard and Thomas F. Hansen contributed equally to this work.

*Correspondence:
Karina Banasik
karina.banasik@cpr.ku.dk
Full list of author information is available at the end of the article

Objective

WGS is becoming increasingly accessible and cost-efficient in both basic and clinical research. Thus, it is relevant to be able to assess whether a given variant exists or whether a given genomic region is constrained or not. Because sequence variants are correlated with geographical location, it is of fundamental importance to have variant counts from different countries and regions when linking phenotype to genotype and many large-scale sequencing studies are for this reason making allele counts available to the research community in an anonymised form (gnomAD etc.). In Denmark, several large studies using genotyping arrays exist, e.g., [1–3], however, few sequencing projects have been conducted and none of them have made the allele counts readily available to the research community. Here, we present DanMAC5, allele counts for sequence variants from 8,671 Danish individuals identified through WGS of three independent studies made available through the accompanying DanMAC5 browser and via EGA. To protect participant privacy and enable a joint data resource, all allele counts below five have been masked and is displayed as <5. The DanMAC5 dataset and browser represents an important open resource of observed single nucleotide variant (SNV) and indel allele counts segregating in the Danish population and can be used for sequence variant filtering in the wider genetics and genomics research community.

Data description

Demographics

Data from three studies were included:

Dan-NICAD: 1,649 individuals with symptoms of obstructive coronary artery disease, predominantly chest pain, undergoing coronary computed tomography angiography. In total, 52% were females, the mean age was 57 years (+/- 9 SD), median coronary artery calcium score were 0 [0–82] and 24% of the cohort had obstructive coronary artery disease defined as >50 diameter stenosis at angiography [4–6].

IBP: Historical data from 3,675 (2,155 females) irrevocably anonymized samples originally collected at the then H:S Sct. Hans Hospital.

Migraine: 3,347 (2,406 females) patients from the Danish Headache Center, including families with clustering of migraine [7, 8].

Permissions for the included studies were obtained from the Danish Data Protection Agency and the appropriate Scientific Ethical Committee system.

Whole genome sequencing

WGS data was generated in three independent research projects [4, 7, 9]. In short, genomic DNA was isolated from frozen whole-blood in EDTA tubes with no DNA amplification or enrichment. Sequencing libraries were prepared using TruSeq PCR-Free (Illumina) and sequenced on the Illumina sequencing platform (NovaSeq 6000 or HiSeq) with S4 flow cells using 2×150 bp paired end sequencing. WGS data underwent quality control using the in-house pipeline at deCODE genetics that has been described previously [10, 11]. Genotype calls were generated per individual with GATK HaplotypeCaller v4.3.3 [12]. The VCF-formatted result files were merged, filtered and aggregate counts generated using bcftools v1.14 [13]. The filtering step was performed as follows: variants with a QUAL-score (QD) < 2.0, Root Mean Square of the mapping quality (MQ) < 40.0, and strand bias by Fisher exact (FS) > 60 were excluded [10]. Anonymized allele counts from each research project were annotated to the GRCh38 version of the human genome (GCA_000001405.15_GRCh38_no_alt_analysis_set.fna [14]) were subsequently merged.

An additional extended quality control was performed by removing low-quality variants using a "whitelist" which was based on a rigid variant calling in two cohorts, Dan-NICAD [4, 5] and migraine [7]; base quality score recalibration (BQSR) was performed using recalibration tables generated with the Sentieon QualCal algorithm. GVCFs were created for each individual using the Haplotyper algorithm before merging with GVCFtyper [15]. Variant quality score recalibration (VQSR) was performed independently for SNPs and indels, based on hapmap3, 1000 genomes, and dbSNP resources, using a sensitivity threshold of 99.7 for passing variants.

After merging and additional quality control filtering using the whitelist, variants with minor allele counts (MAC) of less than five (i.e., seen one to four times) were reported as "<5" to ensure participants' privacy. Sequence variants on the Y chromosome and mitochondria are not reported. A total of 8,671 samples passed the standard quality measures, with an average coverage of 30 reads.

Browser

Using the Dash web-framework (https://plotly.com/dash/) we created an interactive data browser which is available at www.danmac5.dk. Queries can be made using rsID, variant position (chr:pos), gene name (RefSeq), or genomic ranges (chr:pos-pos). All positions are GRCh38/hg38. A hyperlink to gnomAD [16] v3.1.2 (based on hg38) is available in the rsID column. Table 1 lists the file that hold DanMAC5 data and where the features of the DanMAC5 browser are extracted from.

Table 1 Overview of data files/datasets

Label	Name of data file	File types (file extension)	Data repository and identifier
Data file 1	*all_mac5*	*Variant Call Format (.vcf)*	*European Genome-phenome Archive* https://identifiers.org/ega.dataset:EGAD00001009756

Limitations

- Sequence variants cannot be linked to the individual's disease status.
- Our sample contains related individuals which may result in slightly over- or underestimated allele counts.
- Variants with a total allele count below five are listed as < 5 to enable the sharing of data for population genetics and protect the privacy of participants.
- Larger structural variants, variants on the Y chromosome, and mitochondrial variants were not assessed.
- Genomic regions containing repetitive sequences could not be retrieved using pair-end sequencing.

Abbreviations

Dan-NICAD Danish study of Non-Invasive testing in Coronary Artery Disease
EGA European Genome-phenome Archive
GVCF genomic variant call format
IBP Institut for Biologisk Psykiatri (Research Institute of Biological Psychiatry)
MAC minor allele counts
SNV single nucleotide variant
WGS whole-genome sequencing

Acknowledgements

We thank all the participants of the three studies.

Authors' contributions

Draft Manuscript: KB, MN, TW, TFH. Browser: PCH, KB, TFH, SB, MN, DW, PJC, TS. Data analysis: PLM, TRT, AR, AI, PDR, LJAK, MAC, KS, DFG, HS, GBW, ÓÞM, GÁÞ. PIs of included studies: SW, MB, JO, TW, MN, TFH. The author(s) read and approved the final manuscript.

Funding

Open access funding provided by Royal Danish Library. Cost of sequencing was provided through scientific collaboration with deCODE genetics. Karina Banasik, Thomas F. Hansen, and Søren Brunak acknowledge the Novo Nordisk Foundation (NNF17OC0027594 and NNF14CC0001). Simon Winther acknowledges the Novo Nordisk Foundation (NNF21OC0066981). Mette Nyegaard acknowledges the Novo Nordisk Foundation (grant NNF21OC0071050). Thomas F. Hansen and Jes Olesen have received funding from Candy's foundation (CEHEAD).

Availability of data and materials

The DanMAC5 data described in this Data note can be freely and openly accessed on www.danmac5.dk. Please see references [4, 7–9] for details and links to the original studies.

Download

The full dataset is available for academic use for bona fide researchers via https://identifiers.org/ega.dataset:EGAD00001009756 upon registration via the European Genome-phenome Archive: providing clear terms and conditions for use of the full dataset [17]. Please refer to the European Genome-phenome Archive (https://ega-archive.org/access/data-access) for details on how to register.

Declarations

Ethics approval and consent to participate

Permissions for the three independent studies were obtained from the Danish Data Protection Agency and the appropriate Scientific Ethical Committee system (Scientific Ethics Committees of the Central or Capital Region of Denmark) and written consent was obtained from all participants in each study.

Consent for publication

Not applicable.

Competing interests

KB, PLM, TRT, PCH, AI, AR, PDR, LJAK, DW, TS, PJC, MAC, JO, SW, MB, SB, TW, MN and TFH declare no conflict of interest. KS, DFG, HS, GBW, ÓÞM, and GÁÞ are employees of deCODE genetics.

Author details

[1]Translational Disease Systems Biology, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen N, Denmark. [2]Department of Biomedicine, Aarhus University, Høegh-Guldbergsgade 10, DK-8000 Aarhus C, Denmark. [3]Danish Headache Center, Department of Neurology, Copenhagen University Hospital, Valdemar Hansensvej 1-13, DK-2600 Glostrup, Denmark. [4]deCODE genetics, Sturlugata 8, IS-101 Reykjavik, Iceland. [5]Institute for Biological Psychiatry, Mental Health Center Sct Hans, Copenhagen University Hospital, Boeserup vej 2, DK-4000 Roskilde, Denmark. [6]Present Address: Department of Health Science and Technology, Genomic Medicine Group, Aalborg University, Selma Lagerløfs Vej 249, DK-9260 Gistrup, Denmark. [7]Department of Cardiology, University Clinic for Cardiovascular Research, Gødstrup Hospital, Hospitalsvej 15, DK-7400 Herning, Denmark.

Received: 5 August 2022 Accepted: 18 May 2023
Published online: 27 May 2023

References

1. Hansen TF, Banasik K, Erikstrup C, Pedersen OB, Westergaard D, Chmura PJ, et al. DBDS genomic cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. BMJ Open. 2019;9(6):e028401.
2. Laursen IH, Banasik K, Haue AD, Petersen O, Holm PC, Westergaard D, et al. Cohort profile: Copenhagen Hospital Biobank - Cardiovascular Disease Cohort (CHB-CVDC): Construction of a large-scale genetic cohort to facilitate a better understanding of heart diseases. BMJ Open. 2021;11(12):e049709.
3. Pedersen CB, Bybjerg-Grauholm J, Pedersen MG, Grove J, Agerbo E, Bækvad-Hansen M, et al. The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. Mol Psychiatry. 2018;23(1):6–14.
4. Nissen L, Winther S, Isaksen C, Ejlersen JA, Brix L, Urbonaviciene G, et al. Danish study of non-invasive testing in coronary artery disease (Dan-NICAD): study protocol for a randomised controlled trial. Trials. 2016;17:262.

5. Nissen L, Winther S, Westra J, Ejlersen JA, Isaksen C, Rossi A, et al. Diagnosing coronary artery disease after a positive coronary computed tomography angiography: the Dan-NICAD open label, parallel, head to head, randomized controlled diagnostic accuracy trial of cardiovascular magnetic resonance and myocardial perfusion scintigraphy. Eur Heart J Cardiovasc Imaging. 2018;19(4):369–77.
6. Christiansen MK, Nissen L, Winther S, Møller PL, Frost L, Johansen JK et al. Genetic Risk of Coronary Artery Disease, Features of Atherosclerosis, and Coronary Plaque Burden. J Am Heart Assoc. 2020;9(3):e014795.
7. Rasmussen AH, Kogelman LJA, Kristensen DM, Chalmer MA, Olesen J, Hansen TF. Functional gene networks reveal distinct mechanisms segregating in migraine families. Brain. 2020;143(10):2945–56.
8. Chalmer MA, Rasmussen AH, International Headache Genetics Consortium, 23andme Research Team, Kogelman LJA, Olesen J, et al. Chronic migraine: Genetics or environment? Eur J Neurol. 2021;28(5):1726–36.
9. Thygesen JH, Zambach SK, Ingason A, Lundin P, Hansen T, Bertalan M, et al. Linkage and whole genome sequencing identify a locus on 6q25–26 for formal thought disorder and implicate MEF2A regulation. Schizophrenia Research. 2015;169(1):441–6.
10. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. Sci Data. 2017;4(1):170115.
11. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the icelandic population. Nat Genet. 2015;47(5):435–44.
12. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, der Auwera GAV et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv; 2018;201178. [cited 2023 Mar 2] Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/201178v3.
13. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10(2):giab008.
14. GRCh38 reference files. [cited 2023 Mar 2]. Available from: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/.
15. Freed D, Aldana R, Weber JA, Edwards JS. The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. bioRxiv. 2017;115717. [cited 2023 Mar 30] Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/115717v2.
16. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
17. EGA European Genome-Phenome. Archive dataset EGAD00001009756 DanMAC5. [cited 2023 Feb 20]. Available from: https://identifiers.org/ega.dataset:EGAD00001009756.

Publisher's Note