DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors

Hansen, Thomas Folkmann; Banasik, Karina; Erikstrup, Christian; Pedersen, Ole Birger; Westergaard, David; Chmura, Piotr Jaroslaw; Nielsen, Kasper; Thørner, Lise; Hjalgrim, Henrik; Paarup, Helene; Larsen, Margit Anita Hørup; Petersen, Mikkel; Jennum, Poul; Andersen, Steffen; Nyegaard, Mette; Jemec, Gregor Borut Ernst; Olesen, Jes; Werge, Thomas; Johansson, Pår I; Sørensen, Erik; Brunak, Søren; Ullum, Henrik; Burgdorf, Kristoffer Sølvsten

Published in:
BMJ Open

DOI:
10.1136/bmjopen-2018-028401

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors

Thomas Folkmann Hansen,1,2 Karina Banasik,2 Christian Erikstrup,3 Ole Birger Pedersen,4 David Westergaard,2 Piotr Jaroslaw Chmura,2 Kaspar Nielsen,5 Lise Thørner,6 Henrik Hjalgrim,7 Helene Paarup,8 Margit Anita Horup Larsen,6 Mikkel Petersen,3 Poul Jennum,9,10 Steffen Andersen,11 Mette Nyegaard,12 Gregor Borut Ernst Jemec,13 Jes Olesen,1 Thomas Werge,14,15 Pär I Johansson,16 Erik Sørensen,6 Søren Brunak,2 Henrik Ullum,6 Kristoffer Sølvsten Burgdorf6

ABSTRACT

Purpose To establish a cohort that enables identification of genomic factors that influence human health and empower increased blood donor health and safe blood transfusions. Human health is complex and involves several factors, a major one being the genomic aspect. The genomic era has resulted in many consortia encompassing large samples sizes, which has proven successful for identifying genetic factors associated with specific traits. However, it remains a big challenge to establish large cohorts that facilitate studies of the interaction between genetic factors, environmental and life-style factors as these change over the course of life. A major obstacle to such endeavours is that it is difficult to revisit participants to retrieve additional information and obtain longitudinal, consecutive measurements.

Participants Blood donors (n=110000) have given consent to participate in the Danish Blood Donor Study. The study uses the infrastructure of the Danish blood banks.

Findings to date The cohort comprises extensive phenotype data and whole genome genotyping data. Further, it is possible to retrieve additional phenotype data from national registries as well as from the donors at future visits, including consecutive measurements.

Future plans To provide new knowledge on factors influencing our health and thus provide a platform for studying the influence of genomic factors on human health, in particular the interaction between environmental and genetic factors.

INTRODUCTION

A person’s health is determined by complex interactions between genetic, environmental and lifestyle factors. Analysing these factors collectively and prospectively is preferable.

However, this is usually only possible using birth-cohorts and large population-based cohorts, and due to the extensive effort involved in establishing such cohorts, they are rare. The Danish Blood Donor Study (DBDS; dbds.dk) is a large prospective cohort of blood donors aiming at identifying predictors of healthy donors. As part of this cohort, we have now established DBDS Genomic Cohort assessing common single nucleotide polymorphisms (SNPs) in 110000 donors. Thus, the DBDS Genomic Cohort provides a comprehensive catalogue for large-scale genetic analyses in relation to numerous environmental and lifestyle factors affecting donor’s health. A description of summary statistics on phenotypes and data coverage is provided in table 1. A detailed sociodemographic description of Danish blood donors, including sex and age distribution, ethnicity, education, employment and level of urbanisation is found in Burgdorf et al (2017).1

The evaluation of blood donor’s health is important for several reasons. It is crucial for both the donor and the blood recipient that a healthy donor population is maintained.
with a high donation rate and a low dropout rate, thereby ensuring a steady blood supply. Evidence-based guidelines for donor recruitment, care and retention are needed to ensure that donor recruitment can focus on individuals who are likely to remain healthy and donate frequently in the long term. One obvious relevant influential parameter relates to iron metabolism: we know that hundreds of genes impact the generation and regulation of blood cells and also influence phenotype variations of iron absorption and metabolism. Genome-wide SNP information is expected to provide knowledge enabling us to evaluate to whom donating blood will be unproblematic, thus facilitating retention of a stable blood donor population. Another parameter is altruism. Altruism as part of a prosocial behaviour, the selfless concern for the welfare of others, is generally considered a typical blood donor characteristic. However, altruism in the context of voluntary blood donation has also been shown to be a very complex phenotype. In a previous study, we found a substantially larger genetic influence on blood donor behaviour compared with most previous twin studies on altruism, which further highlights the heterogeneity of the blood donor personality. The considerable amount of kinship (twins and siblings) in the cohort in combination with socioeconomic variables from the Danish registries will enable us to further differentiate between genetic impact and social basis of altruism in blood donation. The genome-wide SNP information will provide knowledge that can aid in the identification of long-term and steady donors. Further, we will test for association between genotype SNP information and prodromal symptoms of somatic and psychiatric disease or illnesses.

The DBDS Genomic Cohort offers the possibility to assess the impact of heterogeneous exposures in a broad range of phenotypes, such as mental state, risk-taking behaviour and characterisation of blood components and immune defense. The setup of the study allows for researchers to assess the genetic association in (1) cross sectional studies to investigate, that is, the variation of phenotypic characteristic, clinical and biochemical measurements, (2) retrospective studies of, for example, rehabilitation capacity and (3) prospective studies, for example, analysing the variation of phenotypes and clinical measurements over time and even identify prodromal symptoms.

### COHORT DESCRIPTION

#### Population

The nation-wide blood donor population in Denmark consists of more than 230,000 donors giving more than 300,000 blood donations annually (http://www.bloddonor.dk). Blood donation in Denmark is voluntary and unpaid. This means that the donation is based on the desire to help others, who need it and not the desire of an economic benefit. Blood donors must be physically well, aged 17 and 67 years, and weigh more than 50 kilos. Individuals in chronic medical treatment or frequent travelers to countries with high prevalence of blood disease are not allowed to participate. The deferral rules can be seen at http://www.bloddonor.dk. Blood donors from foreign countries must have lived in Denmark for a minimum of 1 year, have a Danish social security number and have learnt the Danish language to prevent misunderstandings between the donor and the blood bank professionals.

The nationwide Danish blood bank is an integrated part of the Danish healthcare system financed by local and state taxes. The Danish healthcare system is administrated by democratically elected assemblies from national state institutions, regions and municipalities. The Danish blood banks are non-profit organisations owned and operated by each of the five regions in Denmark. The blood banks have a national board to structure collaboration across regions on recruiting donors, processing and distributing the blood for the Danish population. The DBDS is building on the structured Danish blood bank system in the regions responsible for administering donation sites at 27 hospitals in addition to five mobile donation units using 180 selected sites nationally (eg, large companies, sports centres and universities)

At the blood donation centres and attached laboratories, the entire necessary infrastructure needed for the collection of biological samples and structured data is in place. Both blood plasma and whole blood for DNA extraction are available from all donors. The blood bank infrastructure already has laboratory facilities with educated staff (nurses, technicians, IT specialists and physicians). In addition, the blood bank professionals facilitate the testing of the blood for a variety of biomarkers and holds expertise in large scale storage of biological material.

Contact to participants at the blood banks is fundamental to our study. Blood donors are asked to participate and sign an informed consent when they visit the blood bank to donate blood. This consent allows us to use

#### Table 1 Prevalence and missingness of environmental variables

<table>
<thead>
<tr>
<th></th>
<th>Participants (%)</th>
<th>Missingness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>48.8%</td>
<td>0</td>
</tr>
<tr>
<td>Women</td>
<td>51.2%</td>
<td>0</td>
</tr>
<tr>
<td>Age*</td>
<td>41 (29–50)*</td>
<td>0</td>
</tr>
<tr>
<td>Height (cm)*</td>
<td>175.6 (169–182)</td>
<td>0.53</td>
</tr>
<tr>
<td>Weight (kg)*</td>
<td>78.1 (67–87)</td>
<td>0.82</td>
</tr>
<tr>
<td>BMI</td>
<td>25.3 (22–27)</td>
<td>1.2</td>
</tr>
<tr>
<td>Smoking</td>
<td>11.9%</td>
<td>1.3</td>
</tr>
<tr>
<td>Alcohol consumption (daily)</td>
<td>3.0%</td>
<td>4.7</td>
</tr>
<tr>
<td>Questionnaire 1</td>
<td>87 000</td>
<td></td>
</tr>
<tr>
<td>Questionnaire 2</td>
<td>50 000</td>
<td></td>
</tr>
</tbody>
</table>

*Mean (Quartiles).
BMI, body mass index.
the blood samples from their past and future donations to study the impact of genetic and immunological factors on current and future health and disease. The inclusion and exclusion criteria for blood donation and participation in DBDS are the same with 95% of the blood donors who are invited agrees to participate in DBDS.9

**Questionnaires**

From March 2010 until July 2015, all participating donors had to complete a four-page paper-based questionnaire with questions of self-experienced physical and mental health including the 12-item short form (SF-12) standardised health survey, smoking habits, alcohol intake, exercise, food intake, supplemental iron intake, height, weight and waist circumference. Total 85,000 individuals filled out the paper questionnaire. As a follow-up to the initial paper-based questionnaire, we have developed and implemented a digital and flexible tablet-based questionnaire platform, using the open source survey software tool LimeSurvey.10 This enables a rapid, easy and cost-effective procedure to collect self-reported data on health traits from the participating donors at the donor sites at multiple time-points. The first digital questionnaire was implemented and used from July 2015 until May 2018. The questionnaire focused on the following research questions: allergy, ADHD, migraine, hidradenitis, depression and Restless Legs Syndrome. It also contains questions from the paper-based questionnaire: SF-12, smoking habits, alcohol intake, height and weight. In total, 48,000 DBDS participants completed the first digital questionnaire. The second digital questionnaire started in June 2018. It includes questions on: sleep patterns, anxiety, migraine, stress, skin diseases, endometriosis, pain, learning difficulties, SF-12, smoking habits, alcohol intake, height and weight. Using the questionnaire data, several studies have already assessed the factors describing the general health of blood donors, for example, hidradenitis,11 risk of infection,12 migraine, restless legs and depression.1314 Also, by revisiting samples, it has been possible to assess whether infections correlates with obesity15 and how iron deficiency and haemoglobin levels affect donor health in general.1617

**Registry-based data**

There is and has always been equal access to the healthcare system in Denmark; hence, the centralised national civil and health registries are unbiased and comprehensive resources of healthcare data. The registries include the National Patient Registry containing all hospital-registered diagnoses since 197718 as well as other specialised registries, for example, the Danish Medical Birth Register, the Danish Register of Causes of Death and Statistics Denmark monitoring, for example, socioeconomic data. We have already used the DBDS cohort in epidemiological studies assessing, for example, the mortality19 of donors and the effect of blood donation on offspring birth weight.20

**DBDS organisation**

The DBDS itself is described in detail by Pedersen et al.9 Briefly, DBDS is governed by a steering committee with a scientific advisory board. All projects are managed by the DBDS steering committee. Genetic projects involving genetic data in DBDS are run in collaboration between DBDS Genomic Consortium that consists of the DBDS steering committee, deCODE Genetics and scientific collaborators.

Genotyping: DNA is purified from whole blood and subsequent stored at −20°C. All samples are then genotyped in two batches at deCODE genetics using the Global Screening Array by Illumina (batch 1, n=85,000 and batch 2, n=25,000). The array has a very rich up-to-date content of >650,000 SNPs with custom chip content optimised for comparison with the Illumina Omni Express chip. All genotype data are processed simultaneously for genotype calling, quality control and imputation. Initially, individuals or SNPs with more than 10% missing data are excluded, as are individuals deviating more than three SD from the population heterozygosity (correcting for individuals carrying large copy number variations (CNVs) (>100Kbp)).

Imputation

The genotyping data are imputed using a reference panel backbone consisting of (1) UK I KG phase 3 and HapMap reference to predict non-genotyped SNPs with minor allele frequency (MAF)>1% and (2) an in-house dataset consisting of n=6000 Danish whole genome sequences to improve the prediction of variations with a MAF down to around 0.01%.21 Variants listed in the American College of Medical Genetics and Genomics guidelines are currently not predicted, due to restrictions from the Danish National Ethics Committee.22 For future collaborative studies, the Ethics Committee will approve analysis of these variants on a case by case basis.

Copy number variations

Using the genotype of b-allele count and log Ratio, CNVs are called using pennCNV.23 CNVs called using <20SNPs are excluded and the remaining CNVs are visually inspected to exclude false positives.

**Statistical design**

All data are stored and analysed on a specialised, secure section of the 16,000 core Danish National Supercomputer for Life Sciences—Computerome (http://www.computerome.dtu.dk). Data storage and computational analysis is performed on a protected, private cloud environment. The analysis environment is capable of dynamic scaling and has been successfully tested in a composition of over 100 servers totalling more than 300 CPUs, over 13 TB of RAM and has access to up to 5.7 PB (5700 TB) of disk space. The cluster comes with a preconfigured queuing system, possibility to run Virtual Machines and containers (eg, Docker, Singularity), a set of over 900 preinstalled tools and packages and a possibility to add
GPU servers optimised for Machine Learning and specialised big memory systems (1–8 TB of RAM).

For each hypothesis tested, a synopsis is provided including a detailed analysis plan. Information on each synopsis will be published either as link to published articles describing the results or as summary statistic on the study website: http://www.dbds.dk.

**Samples**

During each visit to a blood donor facility (up to four times per year for whole blood donors or up to ten times if plasma donors), every participant donates one EDTA plasma sample. At inclusion in DBDS one whole blood sample is also taken. Plasma samples taken prior to the inclusion date are stored for quality assessments and will also be accessible for future analyses. All samples are frozen within 6 hours of donation and stored in the primary collection tubes until processing.

Routine blood measurements including, for example, blood group, red and white blood cell counts, haemoglobin concentration and haematocrit are obtained at each donation. Besides routine measures, project-related measurements are available, for example, subgroups of patients are assessed for ferritin levels, infection status (Cytomegalovirus, Toxoplasma gondii and Herpes Simplex Virus), HLA-typing and other selected markers of infection (circulating cytokines, C reactive protein and so on).

**General data protection and ethical issues and principles**

DBDS has secured necessary permissions and approval from the Danish Data Protection Agency (2007-58-0015) and the Scientific Ethical Committee system (M-20090237). New projects within the DBDS Genomic consortium will require additional approval by the National Committee on Health Research Ethics. DBDS will be responsible for the continued contact with, and securing future permissions from, relevant Danish authorities regarding research on DBDS samples.

**Patient and public involvement**

Patients and public were not involved in the design of this study.

**Data availability statement**

The study will adhere to the FAIR (http://datafairport.org/: Findable, Accessible, Interoperable and Reusable) concepts. Within this legal framework, DBDS Genomic Consortium Board can thus decide how and under which conditions the data can be shared. Generally, relevant summary data will be publicly available via repositories 3 months after acceptance for publication (H2020 open-access policy).

**FINDINGS TO DATE**

Initial quality measures that have been assessed:

**Kinship**

As described, giving blood often runs in families and the heritability has been estimated to be >53%.8 It is clear from the estimated kinship based on the first batch of participants (n=85,000) (figure 1, table 2), that there is a considerable first, second and third degree relatives among the participants in the DBDS Genomic Cohort.

**Ethnicity**

Based on ~15K overlapping SNPs from the genotyped data and the 1000 Genomes samples, we confirm the expected population structure of the DBDS cohort; most participants are of European ancestry (99%) and the following two ethnicity groups are of South Asian (0.4%) and East Asian (0.2%) ancestry, respectively. The proportion of participants with recent African ancestry is
extremely low (0.002%) which is expected given the strict donor travel quarantine rules. Ethnicity was evaluated using FlashPCA2.24

Minor allele frequencies
The distribution of the MAF shows that majority of SNPs (84%) are above 1% (figure 2) as expected, which provides solid basis for genotype imputation.

STRENGTH AND LIMITATIONS
Consecutive measurements
A unique feature of this large blood donor cohort is the ability to do consecutive assessments. In standard settings, participants are typically recruited at a baseline time-point and are invited for follow-up studies once or twice in the following years. The blood bank represents an advantage because most donors have a long-term committed relationship for blood donation and are seen one to four times annually.25 It is therefore possible to collect several yearly and consecutive biological samples and questionnaire information over decades for a large number of participants. Again, subgroups and samples from specific time-points can be used in a retrospective manner.

National-based registries
Denmark has several comprehensive national registries, which include both health information and sociodemographical measures on an individual level. The informed consent allows for combining information obtained from the DBDS participants and the national registries; the Danish National Patient Registry (since 1977), the Danish Cancer (since 1943) and Diabetes Registries (since 1992–2012), the Danish Registry of Medicinal Product Statistics (since 1994), the Civil Registration Registry (since 1968 vital status, number of children, birthplace, address, relocation and more), the Integrated Database for Labor Market Research (since 1982, eg, educational level, occupation status, income, social status and other related parameters). This facilitates retrospective, cross-sectional or prospective studies using registry-based measurements in combination with questionnaire-derived data.

Donor selection
Although DBDS participants resemble the Danish population, a few limitations in the study design may affect the generalisability of the results.1 The blood donor exclusion criteria dismiss individuals with infections or diseases that are transmittable through blood, weight below 50 kg, haemoglobin (Hb) levels below 12.9 g/L in males and 12.0 g/L in females and curious behaviours: individuals with high travel rates to countries with high risk of hepatitis and HIV, men who have sex with men, individuals who have previously worked as sex workers, those who have used intravenous substances and pregnant women. Comparing sociodemographical parameters of blood donors with that of the total Danish population, we know that very low-income and high-income individuals are underrepresented among blood donors.1 In this respect, we acknowledge that DBDS lacks coverage of certain parts of the general population in contrast to traditional population-based studies. Similarly, the population based UK Biobank study have also reported a ‘healthy volunteer’ bias.26

DISCUSSION
The extension of the DBDS with a genomic cohort will profoundly impact the usability and empower studies on genetic, environmental and lifestyle factors that influence blood donor health. Furthermore, the study provides a unique platform that facilitates analysis of common phenotypes not otherwise found in the national health registries, disease resilience factors and interactions between genes and environment. Finally, such a large healthy cohort holds a huge potential for providing crucial information for future precision medicine initiatives and similar efforts have been started, for example, ‘All of US’ (by NIH, US). In Denmark, we have the advantage of a collection of extensive, national health registries that
facilitate epidemiological studies on specific diseases/outcomes in such a large cohort. For phenotypes and symptoms not monitored systematically in these health registries, for example, lifestyle factors such as smoking habits, sleep patterns and self-perceived health, large epidemiological studies are needed but typically difficult to conduct. The DBDS Genomic Cohort can facilitate such studies. As described above, the DBDS Genomic Cohort exploits an existing blood donor platform with an extremely high participation rate (>95%), which facilitates a straightforward evaluation of donor health in large epidemiological studies. Furthermore, the electronic questionnaire platform allows for easy and fast implementation of new, targeted investigations in subgroups of the donor population. Together with outcomes from the Danish health registries and the millions of retrospective plasma samples stored in easily accessible freezers, DBDS represents a solid phenotyping platform that can be used for both cross-sectional epidemiological studies and for retrospective biomarker studies. We believe that these strengths make the DBDS Genomic Cohort a strong competitive player in the field of precision medicine.

The DBDS Genomic Cohort allows us to study gene-environment interaction that are otherwise difficult to study: testing disease development hypotheses, for example, (1) cognitive performance in interaction with genetic factors and the risk of dementia and (2) determining the contribution of genetic and environmental factors to a phenotype such as sleep pattern. One way to study this is by using multivariate regression models. As an example, Fan et al have recently portrayed an advanced example of gene-environment interaction analysis incorporating temporal and spatial considerations, an analysis based on data from the Danish civil registries. The DBDS Genomic Cohort is particularly valuable for studying disease resistance in individuals exposed to one or more known disease risk factors and yet do not proceed to develop the disease. One such example could be participants carrying a high load of a highly inheritable trait like psychiatric illness in the family or known genetic risk factors, who do not have psychiatric illness themselves. Last, DBDS also provides sequential storage of plasma samples, which allows for sequential blood measurements. Such measurements could be used to investigate health markers like suPAR (soluble urokinase plasminogen activating receptor) and the variation associated with the donor’s general health.

In short, the DBDS Genomic Cohort facilitates the investigation of the impact of genomic factors on health traits and states.

Integrative analysis of different ‘omics, that is, multomics analysis will be possible in the large DBDS Genomic Cohort, which adds tremendously to its value as a resource for studying the health of blood donors and correlation between blood related traits and states. We expect that the DBDS Genomic Cohort will contribute to discovery and validation of prodromal symptoms and biomarkers of disease, thus providing a better understanding of the disease pathologies and suggesting new drug targets.

COLLABORATION
We encourage scientific collaborations using the data generated in the DBDS Genetic Consortium. Published summarised data are available on request. Otherwise request of data necessitates first approval by the DBDS steering committee and if the request is considered outside the aim of DBDS, application to the national scientific ethical committee is obligatory. Additionally, material transfer and data protection agreement need to be acquired. Please visit http://www.dbds.dk.

Author affiliations
1Danish Headache Center, Rigshospitalet Glostrup, Glostrup, Denmark
2Center for Protein Research, Københavns Universitet Sundhedsvidenskabelige Fakultet, København, Denmark
3Immunology, Aarhus Universitetshospitalet, Aarhus, Denmark
4Department of Immunology, Nordhavn Hospital, Næstved, Denmark
5Immunology, Aalborg University Hospital, Aalborg, Denmark
6Immunology, Rigshospitalet, København, Denmark
7Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark
8Immunology, Odense Universitetshospitalet, Odense, Denmark
9Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Glostrup Hospital, Copenhagen, Denmark
10Center for Healthy Ageing, University of Copenhagen, Copenhagen, Denmark
11Economics, Copenhagen Business School, Frederiksberg, Denmark
12Biomedicine, Aarhus Universitet Health, Aarhus, Denmark
13Dermatology, Roskilde University Hospital, Roskilde, Denmark
14Clinical Medicine, Københavns Universitet Sundhedsvidenskabelige Fakultet, København, Denmark
15Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen University Hospital, Roskilde, Denmark
16Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark

Acknowledgements
We wish to express a special thanks to the staffs of the Danish blood banks whose continued inclusion of blood-donors into the DBDS makes this research possible.

Contributors TFH, KB and KSB conceived and planned the experiments. KB carried out the analyses. OBP, HH, HP, KN, CE, HI, PJ, TW, JD, OBEJ, MN, SA, PLU, JS and LT contributed to cohort and research design. DW, PJC, KB and SB led to data infrastructure design. CE, KSB, MAHL, ES and MP contributed to data capture. TFH, KB and KSB contributed to the interpretation of the results. TFH, KB and KSB took the lead in writing the manuscript. All authors provided critical feedback and helped shape the analysis and manuscript and approved the final version.

Funding We wish to thank the Lundbeck Foundation, Denmark under Grant (R209-2015-3500 to Kristoffer Burgdorf), The Danish Administrative Regions, The Danish Blood Donor Research Foundation (Bloeddonorernes Forskningsfond), Rigshospitalets Research Foundation, The Novo Nordisk Foundation (NNF14CC0001 and NNF1700027594) and CANDY foundation (CENHEAD).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed. See data sharing statement.

Data sharing statement The study will adhere to the FAIR (http://datafairport.org/): Findable, Accessible, Interoperable and Reusable) concepts. Within this legal framework DBDS Genetic Consortium Board can thus decide how and under which conditions the data can be shared. Generally, relevant summarydata will be publicly available via repositories 3 months after acceptance forpublication (H2020 open-access policy).

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

REFERENCES