Molecular Genetics and Metabolism 110 (2013) 65–72

Contents lists available at ScienceDirect

Molecular Genetics and Metabolism

journal homepage: www.elsevier.com/locate/ymgme

Archived neonatal dried blood spot samples can be used for accurate whole genome and exome-targeted next-generation sequencing

Mads Vilhelm Hollegaard [a,*], Jonas Grauholm [b], Ronni Nielsen [c], Jakob Grove [d,e], Susanne Mandrup [c], David Michael Hougaard [a]

[a] Department of Clinical Biochemistry, Immunology and Genetics, Statens Serum Institut, Artillerivej, Copenhagen, Denmark
[b] AROS Applied Biotechnology A/S, Science Park Skejby, Brendstrupgaardsvej, Aarhus, Denmark
[c] Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej, Odense, Denmark
[d] Department of BiomedicineBartholins Allé, Aarhus University, Aarhus, Denmark
[e] Bioinformatics Research Centre (BiRC), C.F. Møllers Allé, Aarhus University, Aarhus, Denmark

ARTICLE INFO

Article history:
Received 11 May 2013
Received in revised form 3 June 2013
Accepted 4 June 2013
Available online 13 June 2013

Keywords:
Neonatal dried blood spot samples
Guthrie cards
Next-generation sequencing
Whole-genome sequencing
Whole-exome sequencing

ABSTRACT

Dried blood spot samples (DBSS) have been collected and stored for decades as part of newborn screening programmes worldwide. Representing almost an entire population under a certain age and collected with virtually no bias, the Newborn Screening Biobanks are of immense value in medical studies, for example, to examine the genetics of various disorders. We have previously demonstrated that DNA extracted from a fraction (2 × 3.2 mm discs) of an archived DBSS can be whole genome amplified (wgaDNA) and used for accurate array genotyping. However, until now, it has been uncertain whether wgaDNA from DBSS can be used for accurate whole genome sequencing (WGS) and exome sequencing (WES).

This study examined two individuals represented by three different types of samples each: whole-blood (reference samples), 3-year-old DBSS spotted with reference material (refDBSS), and 27- to 29-year-old archived neonatal DBSS (neoDBSS) stored at −20 °C in the Danish Newborn Screening Biobank. The reference samples were genotyped using an Illumina Omni2.5M array, and all samples were sequenced on a HighSeq2000 Paired-End flow cell. First, we compared the array single nucleotide polymorphism (SNP) genotype data to the single nucleotide variation (SNV) calls from the WGS and WES SNV calls. We also compared the WGS and WES reference sample SNV calls to the DBSS SNV calls.

The overall performance of the archived DBSS was similar to the whole blood reference sample. Plotting the error rates relative to coverage revealed that the error rates of DBSS were similar to that of their reference samples. SNVs called with a coverage $< \times 8$ had error rates between 1.5 and 35%, whereas the error rates of SNVs called with a coverage ≥ 8 were <1.5%. In conclusion, the wgaDNA amplified from both new and old neonatal DBSS perform as well as their whole-blood reference samples with regards to error rates, strongly indicating that neonatal DBSS collected shortly after birth and stored for decades comprise an excellent resource for NGS studies of disease.

1. Introduction

The opportunity to perform extensive genotyping on DNA extracted from dried blood spot samples (DBSS) used in the newborn screening programmes has opened new avenues in newborn screening as well as for the study of the genetic influence of many complex disorders [1–4]. Residual newborn DBSS stored in national repositories combined with relevant clinical information from national medical registries provide access to large cohorts of well-characterised patients and healthy controls [5]. However, the major challenge using the DBSS for such ventures is the small amount of blood available. In theory, the amount of genomic DNA

Abbreviations: DBSS, dried blood spot samples; gDNA, genomic DNA; wgaDNA, whole-genome amplified DNA; DNSB, Danish Newborn Screening Biobank; NGS, next generation sequencing; WGS, whole-genome sequencing; WES, whole exome sequencing; SNV, single nucleotide variation; SNP, single nucleotide polymorphism; Ref, adult whole blood gDNA reference sample; refDBSS, adult whole blood reference sample stored as a DBSS; neoDBSS, neonatal heel-prick whole-blood sample; E1 error-rate, one basepair discrepancy in SNV/SNP genotype calls; E2 error-rate, two basepairs discrepancy in SNV/SNP genotype calls; bp, base pair.

* Corresponding author.
E-mail addresses: mvh@ssi.dk (M.V. Hollegaard), jg@arosab.com (J. Grauholm), ronni.n@bmb.sdu.dk (R. Nielsen), grove@hum-gen.au.dk (J. Grove), s.mandrup@bmb.sdu.dk (S. Mandrup), dh@ssi.dk (D.M. Hougaard).

1096-7192/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymgme.2013.06.004

(gDNA) that can be extracted from a 3.2 mm punch of DBSS is approximately 60 ng [6]. This shortage of DNA may be overcome by whole genome amplified (wga) DNA (wgaDNA). wgaDNA amplified from DBSS stored at −20 °C/−4 °F for three decades in the Danish Newborn Screening Biobank (DNSB) has been shown to provide highly reliable genotyping results by different approaches: Sanger sequencing [7], genome wide arraying [1,8], high-resolution melting curve analysis [7], and simple SNP genotyping [3]. The excess DBSS material from the Danish Newborn Screening programme has been stored in the DNSB [5] since 1982 and covers close to 100% of the Danish population born since that time. Combined with data from numerous Danish registries, the DBSS has been used successfully in several genetic studies of disease looking at cerebral palsy [9], schizophrenia [10–16], birth weight [17], psychosis [18,19], and infantile hypertrophic pyloric stenosis [20]. In many countries and states, the neonatal screening programmes are storing the excess DBSS in biobanks, if allowed by law. In contrast to the DNSB and the California Research-Ready Biospecimen Bank, the majority of these biobanks, for example, the Michigan neonatal biobank, and the Swedish PKU-Biobank, store their DBSS at room temperature and not at −20 °C/−4 °F, which can complicate their use in some approaches due to degradation of the DNA. However, the potential of using the millions of already stored neonatal DBSS for research on different early onset diseases is enormous and is likely to improve the screening of newborns in the future, and to increase our biological knowledge about many diseases.

Next-generation sequencing (NGS) is currently a rapidly advancing technology within the field of genetics and is moving towards reasonable prices per sample in increasingly fast increments. Coupled with advances in data handling and analysis, this technology is on a path to becoming a standard tool in research and clinical genetics. In addition, this sequencing technology has prodigious potential for disease diagnostics and in the screening of newborns. The sensitivity of the technology is rapidly improving, with current standardised off the shelf protocols allowing the use of nanograms of high-quality input DNA for whole-genome sequencing (WGS), although larger amounts are still necessary for targeted sequencing approaches such as whole-exome sequencing (WES). A solution to the low concentrations of DNA obtainable from DBSS could be whole-genome amplification; however, it is uncertain whether the amplification process compromises the results and introduces genotype call errors or more serious de-novo mutations.

This proof-of-principle study investigates whether it is possible to use only two 3.2 mm discs from archived DBSS for DNA extraction, WGA, and subsequent WES and WGS, using standard protocols. Moreover, by comparing different sample types we aim to evaluate the crude error rates of the WES and WGS single nucleotide variation (SNV) calls by simple comparisons of called variations without using any data optimising steps or algorithms.

2. Materials and methods

2.1. Study overview and sample preparations

The study included two individuals (siblings); A (male) and B (female). As defined by the “Danish Act on Research Ethics Review of Health Research Projects” Section 2, this project does not constitute a health research project but is considered a quality developmental project for newborn screening. The project can thus be conducted without approval from the Committees on Biomedical Research Ethics for the Capital Region of Denmark.

From each of the two adult individuals, a venous blood sample was drawn in a heparin collection tube (Reference (Ref.) sample). For a study overview, please see Fig. 1. A 50 μL fraction of this sample was transferred to Whatman Specimen 903® Collection Paper, air-dried for 3 h at room temperature and thereafter stored for over three years at −20 °C/−4 °F (refDBSS) in the DNSB. Finally, the neonatal DBSS (neoDBSS) from each individual was isolated from the DNSB. DNA was extracted from the Ref. sample using the Maxwell® 16 LEV Blood DNA Kit (Promega) according to the manufacturer's instructions. Two 3.2 mm discs were punched from each of the refDBSS and neoDBSS and DNA was extracted with Extract-N-Amp Blood PCR Kit (Sigma) as previously described [1]. To reach the necessary amounts of DNA for NGS, the DBSS-extracted DNA was WGA in triplicate with REPLIg (QIAGEN) as previously described [1]. The concentration of the Ref. sample, refDBSS and neoDBSS DNA samples was estimated using Quant-IT PicoGreen dsDNA Reagent (Invitrogen).

2.2. Array genotyping and sequencing

As a “gold-standard” reference for the SNVs identified by NGS, we chose to array genotype the Ref. samples using the HumanOmni2.5-Quad BeadChip array (Illumina Inc.). This array covers approximately 2.5 million single nucleotide polymorphisms (SNP) located with a median space between polymorphisms of 0.63 kb. According to Illumina, the call frequency is expected to be above 99% with a reproducibility rate above 99.9%. For this array, we used 200 ng of Ref. sample DNA. Data were loaded into GenomeStudio version 2011.1 using the HumanOmni2.5-4v1_H manifest and the Illumina standard cluster position file for this array.

The libraries for WES were created from 1 μg of double stranded DNA using the TruSeq DNA library prep kit v2 (Illumina). The libraries were quantified using the KAPA Library Quant Kit (KAPAbiosystems) and 500 ng of each library was pooled from six samples and enriched using the TruSeq Exome enrichment kit (Illumina). After targeting, each pool was quantified with the KAPA Library Quant kit (KAPAbiosystems) and sequenced using two lanes of a HiSeq2000 paired end flow cell and a read length of 101 bases (Illumina).

The libraries for WGS were created from 50 ng of double stranded DNA using the Nextera DNA sample prep kit (Illumina). Each sample was quantified using the KAPA Library Quant kit (KAPAbiosystems) and sequenced on a single lane of a HiSeq2000 paired end flow cell with a read length of 101 bases.

2.3. NGS alignment and variant calling

Both the WES and WGS genome data alignments were done using the tools included in the CASAVA 1.8.2 package from Illumina using the first 100 generated bases.

For the WGS data, no de-multiplexing was required as only one sample was sequenced per lane. For the WGS dataset, all reads from a given lane was assigned to the corresponding sample in the Bcl conversion process. The exome pools were multiplexed and de-multiplexing was carried out along with the Bcl conversion while allowing for one mistake in a given index; each used index has at least two differences compared to the other used indices.

Both of the datasets were aligned using CASAVA's Efficient Large-Scale Alignment of Nucleotide Databases v2e (ELANDv2e) and the reference genome used was the Ensemble GRCh37 reference available from iGenomes (Illumina).

The WGS variant calling was accomplished using the CASAVA “configureBuild” command. This command handles sorting of reads and removal of duplicates before calling SNVs and Indels. The variant calling also produced overview statistics used to compare across samples. For the WES dataset, variant detection was performed using the TruSeq Exome Run Quality Script v1.1. This script configures variant calling as with the “configureBuild” command also used for WGS data, additionally it also extract those variants within the target.