Molecular Genetics and Metabolism Reports 10 (2017) 33–37

Contents lists available at ScienceDirect

Molecular Genetics and Metabolism Reports

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

RNA sequencing of archived neonatal dried blood spots

Jonas Bybjerg-Grauholm [a,b,*,1], Christian Munch Hagen [a,b,1], Sok Kean Khoo [c], Maria Louise Johannesen [a,b], Christine Søholm Hansen [a,b], Marie Bækvad-Hansen [a,b], Michael Christiansen [a,b,d], David Michael Hougaard [a,b], Mads V. Hollegaard [a,b]

[a] Department of Congenital Disorders, Statens Serum Institut, Copenhagen DK-2300, Denmark
[b] iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Denmark
[c] Department of Cell and Molecular Biology, Grand Valley State University, Grand Rapids, MI 49503, USA
[d] Department of Biomedicine, University of Copenhagen, Copenhagen N DK-2200, Denmark

ARTICLE INFO

Article history:
Received 25 October 2016
Received in revised form 16 December 2016
Accepted 16 December 2016
Available online 24 December 2016

Keywords:
Gene expression
RNA-seq
Limited material
Neonatal Screening
Guthrie Cards
Dried blood spots

ABSTRACT

Neonatal dried blood spots (DBS) are routinely collected on standard Guthrie cards for all-comprising national newborn screening programs for inborn errors of metabolism, hypothyroidism and other diseases. In Denmark, the Guthrie cards are stored at −20 °C in the Danish Neonatal Screening Biobank and each sample is linked to elaborate social and medical registries. This provides a unique biospecimen repository to enable large population research at a perinatal level. Here, we demonstrate the feasibility to obtain gene expression data from DBS using next-generation RNA sequencing (RNA-seq). RNA-seq was performed on five males and five females. Sequencing results have an average of >30 million reads per sample. 26,799 annotated features can be identified with 64% features detectable without fragments per kilobase of transcript per million mapped reads (FPKM) cutoff; number of detectable features dropped to 18% when FPKM ≥ 1. Sex can be discriminated using blood-based sex-specific gene set identified by the Genotype-Tissue Expression consortium. Here, we demonstrate the feasibility to acquire biologically-relevant gene expression from DBS using RNA-seq which provide a new avenue to investigate perinatal diseases in a high throughput manner.

1. Introduction

Gene expression analysis captures a snapshot of cellular activity that reflects the response to genetic, environmental and epigenetic changes in a biological system. Determining phenotypes at the molecular level with differentially-expressed genes, using high-throughput technologies such as microarray and next-generation RNA-sequencing (RNA-seq), could help bridge the current gap between genotypes and phenotypes [1]. Neonatal dried blood spots (DBS) represent a unique biospecimen resource that can provide the "baseline" gene expression at birth for genotype-phenotype studies. In many developed countries, DBS are routinely collected on standard Guthrie cards for newborn screening to diagnose inborn errors of metabolism, hypothyroidism and other diseases [2]. Post-diagnostic gene expression analysis of DBS can potentially give insights into the cellular response of molecular pathological factors and thereby the etiology of a specific disease. For example, it has been shown that microarray gene expression profiles from DBS can predict children with cerebral palsy [3].

Denmark has an all-comprising comprehensive system of registries which hold and collect information on social and health related life events for each resident linked through a unique person identifiable number (CPR number) [4–6]. Since the early 1980s, the Danish Neonatal Screening Biobank (DNSB) has been systematically collecting and storing surplus DBS from the neonatal screening program. While the primary purpose of DBS is to diagnose and treat congenital disorders, it can also be used for research purposes with appropriate approval [7,8]. Several methods have been developed, primarily concerned with utilizing DNA from DBS [9–12]. In addition, we and several other groups have shown the feasibility of obtaining RNA and generating RNA gene expression microarray data from DBS for various studies [3,13–19].

RNA microarray technologies are fast, robust and relatively inexpensive. However, they have inherent problems such as differing hybridization properties of probes, as well as predominant probe placements near the 3′ end of transcripts. RNA-seq, a cutting-edge massively parallel next-generation sequencing technique, has the capacity to sequence each transcript on a single nucleotide basis. While above median transcripts are generally reliably detected by both RNA microarrays and RNA-seq, RNA-seq offers several significant advantages over microarray

Abbreviations: DBS, Dried blood spots; DNSB, Danish Neonatal Screening Biobank; GTEx, Genotype-Tissue Expression consortium.

* Corresponding author at: Artillerivej 5 85#240, DK-2300 Copenhagen S, Denmark.
E-mail address: jogr@ssi.dk (J. Bybjerg-Grauholm).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymgmr.2016.12.004

such as the capability to distinguish isoforms, call sequence variants, and provide improved accuracy for low-abundance transcripts [20–22]. However, the nature of total RNA, from blood in particular, does complicate sequencing approaches. Firstly, approximately 80% of total RNA from any tissue source is ribosomal RNA (rRNA). Thus, rRNA is usually reduced or removed prior to RNA-seq [23,24]. Secondly, in blood, 50–90% of mRNA transcripts are globin species [21,25]; globin may contribute to uninformative reads that are present in high abundance thus compromising RNA-seq analysis. Globin depletion can remove ~80% of globin species, improve correlation of technical replicates, and enable an additional 3500 RNA transcripts to be detected [21]. In neonatal DBS, globin is primarily derived from fetal hemoglobin (HbF), the predominant form of hemoglobin in the developing fetus which persists in newborn blood until gradually switching over to adult hemoglobin (HbA) ~ 6 months post-birth [26,27]. HbF could be depleted using the same methodology as HbA, but it remains untested whether sequence similarity is high enough to use the exact same method.

To date, RNA-seq for neonatal DBS has not been reported. Here, we conducted a proof-of-concept study to demonstrate the feasibility of obtaining robust and biologically-relevant gene expression data from DBS using RNA-seq.

2. Methods

2.1. DBS samples

This study does not constitute a health-related research project as defined by the Danish "Act on Research Ethics Review of Health Research Projects". It is considered a developmental project for the Newborn Screening Program instead which does not require a separate approval from the Committees on Biomedical Research Ethics for the Capital Region of Denmark. In this study, we used a set of anonymized neonatal DBS samples that had been stripped of all information including the year of sampling. Sexes were determined using Sequenom's Sample ID panel (data not included). Sample names are PKU_M_0 to PKU_M_4 for five male DBS and PKU_F_0 to PKU_F_4 for five female DBS.

2.2. RNA isolation, library preparation and sequencing

Total RNA from two 3.2 mm DBS punches was extracted, purified and concentrated using Illustra RNAspin mini (GE Healthcare) and then subjected to DNase treatment to eliminate DNA contamination. Samples were concentrated using RNA Clean & Concentrator (Zymo Research). Sequencing libraries were prepared according to the standard protocol for the Stranded TruSeq RiboZero-Globin kit (Illumina) which includes Globin and rRNA depletion. RNA was fragmented for 4 min at 94 °C and each library was quantified using the KAPA library quantification kit (KAPA). Libraries were pooled before sequencing with a NextSeq-500 sequencer using NextSeq Control Software v1.3.0 (Illumina) on a high output flow cell (v1) with paired end reads of 76 bp (Illumina). Following sequencing, data was de-multiplexed and converted to FASTQ using bcl2fastq v2.17.1.14 (Illumina).

2.3. Data analysis

The FASTQ files were processed with the settings described by Trepnell [28] using Tophat v2.0.13 and Cufflinks v2.2.5. Filtering of genes and visualization of results was performed in cummerBund 2.8.2 [28] using the commands described by Trepnell [28] within the supplementary information for that paper. The reference genome and gene annotations used was UCSC HG19 prepackaged and downloaded via iGenomes (Illumina, downloaded June 2014).

3. Results

3.1. Library preparation and sequencing statistics

More than 325 million reads were generated from the 10 samples. The average reads per sample was 32.7 million ± 18.3 million, with 34.4 million ± 18.7 million for females and 31.1 million ± 20.0 million for males. The read length was 32–75 bp with an average of 74 bp, 91.5% of bases exceeded Q30. Detailed information of each sample's preparation and sequencing is shown in Table 1. Variation was observed in the dynamic range of transcript intensities, i.e. fragments per kilobase of transcript per million mapped reads (FPKM). Variance in FPKM was more profound across samples than between the sexes (Supplementary Fig. 1).

3.2. Transcript detection

To evaluate transcript detection rates, we limit the read alignment to redundant annotated features (transcripts and non-coding RNA). In total, our reference contained 26,799 features. An average of 64% of all features were detected (SD = 15%, range 32–81%) without FPKM cutoff while the number of detectable features dropped to 18% (SD = 7%, range 3–26%) when FPKM ≥ 1 was applied.

To test for potential outliers, dimensional reduction analysis with Multi-Dimensional Scaling (MDS) and Principle Component Analysis (PCA) were used. Both MDS and PCA exhibited two same potential outliers PKU_M_3 and PKU_F_1, with MDS showed a potential third outlier PKU_M_0 (Supplementary Fig. 2). We did not exclude the outliers from subsequent analysis of differential expression analysis based on sexes.

3.3. Differentiation of sex-specific gene expression

To verify whether DBS can reveal biologically-relevant gene expression using RNA-seq, we evaluated a gene set (19 sex chromosome-linked genes; 21 features) identified by the Genotype-Tissue Expression (GTEx) consortium [29] that are differentially-expressed between male and female in blood. Hierarchical clustering from FPKM showed complete separation of males from females (Fig. 1A). Gene expression heat map showed female samples expressed only X chromosome-linked genes (*XIST* and *TXLNG*) and none of Y chromosome-linked genes while male samples expressed >50% of Y chromosome-linked genes (Fig. 1B). FPKM values for each sex-specific gene are shown in Supplementary Table 1.

3.4. Effect of globin depletion

In this study, although the exact efficiency of globin depletion is unquantifiable, our results showed that globin did not represent the majority of transcripts. To access the effect of globin depletion on RNA-seq data, FPKM values of eight hemoglobin transcripts (*HBBP1*, *HBG1*, *HBG2*, *HBE1*, *HBZ*, *HBM*, *HBA1* and *HBQ1*) were used to create hierarchical clustering and heatmaps. PKU_M_1 has higher FPKM values for *HBG1* and *HBG2* when compared with the rest of the samples. Interestingly, by using the entire hemoglobin gene set in the analysis, female samples can be discriminated from male (Fig. 2A and B). However, when *HBA1* is removed from the analysis, sexes of the samples could not be differentiated (Fig. 2C and D). Analysis using the full set of transcripts separated the three previously identified outliers into another cluster (Fig. 2E). FPKM values for hemoglobin transcripts are shown in Supplementary Table 2.

4. Discussion

Our results show that we can perform RNA-seq on archived DBS samples, despite having some outliers in the dimensional reduction analysis. Sufficient data can be generated for hypothesis driven research