ARTICLE OPEN

Scalable, high quality, whole genome sequencing from archived, newborn, dried blood spots

Yan Ding[1,5], Mallory Owen[1,5]✉, Jennie Le[1], Sergey Batalov[1], Kevin Chau[1], Yong Hyun Kwon[1], Lucita Van Der Kraan[1], Zaira Bezares-Orin[1], Zhanyang Zhu[1], Narayanan Veeraraghavan[1], Shareef Nahas[1], Matthew Bainbridge[1], Joe Gleeson[1,2], Rebecca J. Baer[2,3], Gretchen Bandoli[2], Christina Chambers[2] and Stephen F. Kingsmore[1,4]✉

Universal newborn screening (NBS) is a highly successful public health intervention. Archived dried bloodspots (DBS) collected for NBS represent a rich resource for population genomic studies. To fully harness this resource in such studies, DBS must yield high-quality genomic DNA (gDNA) for whole genome sequencing (WGS). In this pilot study, we hypothesized that gDNA of sufficient quality and quantity for WGS could be extracted from archived DBS up to 20 years old without PCR (Polymerase Chain Reaction) amplification. We describe simple methods for gDNA extraction and WGS library preparation from several types of DBS. We tested these methods in DBS from 25 individuals who had previously undergone diagnostic, clinical WGS and 29 randomly selected DBS cards collected for NBS from the California State Biobank. While gDNA from DBS had significantly less yield than from EDTA blood from the same individuals, it was of sufficient quality and quantity for WGS without PCR. All samples DBS yielded WGS that met quality control metrics for high-confidence variant calling. Twenty-eight variants of various types that had been reported clinically in 19 samples were recapitulated in WGS from DBS. There were no significant effects of age or paper type on WGS quality. Archived DBS appear to be a suitable sample type for WGS in population genomic studies.

npj Genomic Medicine (2023)8:5 ; https://doi.org/10.1038/s41525-023-00349-w

INTRODUCTION

Newborn dried blood spots (DBS) are used worldwide to screen for childhood genetic diseases with effective treatments. Over the past 50 years, universal newborn screening (NBS) has proven an incredibly successful public health intervention for reducing morbidity and mortality due to certain selected conditions[1–8]. Archived DBS represent the largest repository of human genetic material in existence[2,5,9,10]. In the United States, approximately 4 million newborns are screened each year, and some states store DBS from these infants. The California Biobank Program represents the combined biospecimen and data resources of the California Genetic Disease Screening Program and the California Birth Defects Monitoring Program. These programs began tracking birth defects in 1983, and currently screen approximately 400,000 newborns each year for treatable, infant onset diseases. Currently, the California Recommended Uniform Screening Panel includes 33 primary disorders and 51 secondary disorders[2,9,11]. Newborn screening is currently undertaken at California Department of Health Services screening laboratories, and consists of specific assays for each set of disorders. Currently, DNA sequencing is not part of the primary screen, and is only undertaken as a confirmatory test in some states after a primary screen returns an abnormal result.

DBS represents an invaluable resource for research aimed at elucidating the underlying etiology of human diseases, including birth defects, metabolic disorders, and congenital heart defects, and may be particularly suitable for newly available methods for whole genome sequencing (WGS)[1,6,12,13]. Additionally, for infants who die shortly after birth, DBS may be resourced for investigation of the molecular cause of death[14–16]. DBS have numerous advantages over whole blood samples, including ease of transport, potential for storage at room temperature, cost effectiveness, and feasibility of long-term storage without degradation of DNA[10,17–20]. Despite the success of newborn screening, a significant number of treatable genetic diseases go undiagnosed at birth[14–16]. The cost and time required for WGS has rapidly decreased over the past decade[21–24]. In the future, newborn screening for genetic diseases with effective treatments could be greatly expanded by WGS of DBS.

Several issues have been previously identified when using archived DBS for WGS. It has been suggested that the age of the DBS, storage conditions, and filter paper type may impact the yield and quality of the sequencing data[10,17–20,25–27]. Previous studies have shown mixed results when examining these potential sources of variability. For example, Hollegaard et al. found a significant effect of storage length on quality of extracted DNA in samples dating back to 1981[17]. However, the sequencing data still demonstrated an average call rate of 97% for single nucleotide variants (SNVs)[12]. This result was replicated in a recent study by Sok et al.[27] Most studies have examined whole-exome sequencing (WES) from DBS, not WGS. More recently, Bassaganyas et al. reported using DNA isolated from archived DBS to perform WES and WGS with limited polymerase chain reaction (PCR) cycling post-ligation and concluded that the DBS were a satisfactory source of high quality DNA[20]. Of note, none of these studies evaluated structural variant calls. Finally, previous studies specifically assessing WGS from DBS have included small sample sizes and the gDNA yield has been low.

WGS using archived DBS samples from publicly held biorepositories presents the potential to investigate genetic disease at a population level. In this pilot study, we hypothesized that using the newest techniques for WGS without PCR amplification would result in high-quality sequence irrespective of the DBS age or filter paper type. We describe simple, scalable methods for WGS on

[1]Rady Children's Institute for Genomic Medicine, Rady Children's Hospital, San Diego, CA 92123, USA. [2]Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA. [3]California Preterm Birth Initiative, University of California San Francisco, San Francisco, CA, USA. [4]Keck Graduate Institute, Claremont, CA 91711, USA. [5]These authors contributed equally: Yan Ding, Mallory Owen. ✉email: mowen1@rchsd.org; skingsmore@rchsd.org

Table 1. Concordance of variants reported clinically in diagnostic WGS of 25 blood samples and WGS from 63 manufactured DBS.

Subject ID	Relationship to proband	Variants previously reported following diagnostic WGS	Variant classification	FTA DBS WGS	PC DBS WGS	Dx confirmed[a]
1	Proband	*FOXP3*, c.1010G>A, p.Arg337Gln	P	2	3	Y
2	Proband	None	None	1	2	
3	Proband	Chr22:23961084-24401339del	P	1	2	Y
4	Father	None	None	2	2	
5	Sibling	*FLNA*, c.2410G>A, p.Val804Ile	VUS	1	2	Y
6	Proband	*FLNA*, c.2410G>A p.Val804Ile; Chr2:112658998-112854380dup	VUS	1	2	Y
7	Proband	*DCLRE1C*, c.406G>A, p.Asp136Asn; Chr10:14983601-15065700del	LP, P	1	2	Y
8	Proband	*RYR2*, c.12290A>G, p.Asn4097Ser; *ANK2*,c.8404G>C, p.Asp2802His	VUS	1	1	Y
9	Father	*RYR2*, c.12290A>G, p.Asn4097Ser; *ANK2*,c.8404G>C, p.Asp2802His	VUS	1	1	Y
10	Proband	*SMN1* Chr5:70247540-70247820x0 del; *SMN2* Chr5:69372123-69372400dup	P	2	3	Y
11	Proband	*TUBB3*, c.1228G>A, p.Glu410Lys; Chr22:18873501-21466000del	P	3	3	Y
12	Proband	*ROBO1*, c.107G>T, p.Arg36Met; *ROBO1*, c.4610G>A, p.Gly1537Glu	VUS	1	2	Y
13	Proband	*PROKR2*, c.563C>T, p.Ser188Leu; Chr14:59001701-61049600dup	VUS	2	2	Y
14	Proband	*HSD17B4*, c.1619A>G, p.His540Arg	VUS	1	2	Y
15	Proband	None	None	1	1	
16	Proband	None	None		1	
17	Proband	None	None		1	
18	Proband	*EPG5*, c.2066del, p.Leu689Ter	P	1	1	Y
19	Proband	*EPG5*, c.2066del, p.Leu689Ter	P	1	1	Y
20	Proband	None	None	1		
21	Proband	Chr15:23512201-28700800del	P	1		Y
22	Proband	Chr10:81634801-89151100 del	P	1		Y
23	Proband	Chr2:21240919-21244369del	P	1		Y
24	Proband	ChrX:1422154-1423912del	P	1		Y
25	Proband	*SMN1* Chr5:70247540-70247820x0del; *SMN2* Chr5:69372122-69372400x2 dup	P	1		Y

P pathogenic, *VUS* variant of uncertain significance, *LP* likely pathogenic, *del* deletion, *dup* duplication.
[a]Confirmed diagnosis using standard annotation, variant alignment, and analysis pipelines as detailed in "Methods".

laboratory created DBS and randomly selected, de-identified DBS from the California Biobank Program.

RESULTS

Quality and quantity of genomic DNA extracted from manufactured DBS

We manufactured 63 DBS from EDTA-blood samples remnants from 25 individuals who had previously received diagnostic PCR-free WGS on Illumina Novaseq 6000 instruments (Supplementary Table 1). In 19 individuals, at least one variant had been reported clinically (Table 1). DBS were made with two types of filter paper that are widely used for NBS (FTA, ThermoFisher and PC, GE Healthcare; Table 1). Genomic DNA (gDNA) was extracted without noticeable degradation from both types of DBS using two different sample preparation methods (Illumina PCR-free genomic library preparation [Illumina] and KAPA HyperPlus [QIAGEN]; Fig. 1; Supplementary Table 1). The QIAGEN isolation method generated more gDNA per extraction than Illumina (average 868 and 349 ng, respectively, $p < 0.01$, minimum 369 ng and 165 ng, respectively; Supplementary Table 2). gDNA from DBS had a slightly smaller molecular weight than that isolated from the corresponding EDTA-blood but was acceptable for library construction (Fig. 1). The A260/A280 ratio of gDNA generated with QIAGEN was higher than that of Illumina (average 1.72 and 1.57, respectively, $p < 0.01$; Supplementary Table 2). In comparison, 350 µL EDTA-blood from the same individuals yielded an average of 6884 ng gDNA with an A260/A280 ratio of 1.81 (Supplementary Table 2). The filter paper type (FTA or PC DBS) did not affect the concentration of gDNA extracted nor the A260/A280 ratio ($p > 0.05$).

Titration of quantity of genomic DNA extracted from DBS punches

A significant issue in using archived DBS is the number of punches available. Often, DBS have undergone prior extractions for NBS or research purposes, and therefore the number of punches available is decreased. We performed titration experiments with six DBS to evaluate the minimal number of 3 mm diameter punches required for WGS. The median gDNA isolated from 1, 3, 5, 7, and 10 punches was 53, 123, 216, 421, and 707 ng, respectively (Fig. 2a). Thus, six DBS punches were required to isolate at least 200 ng

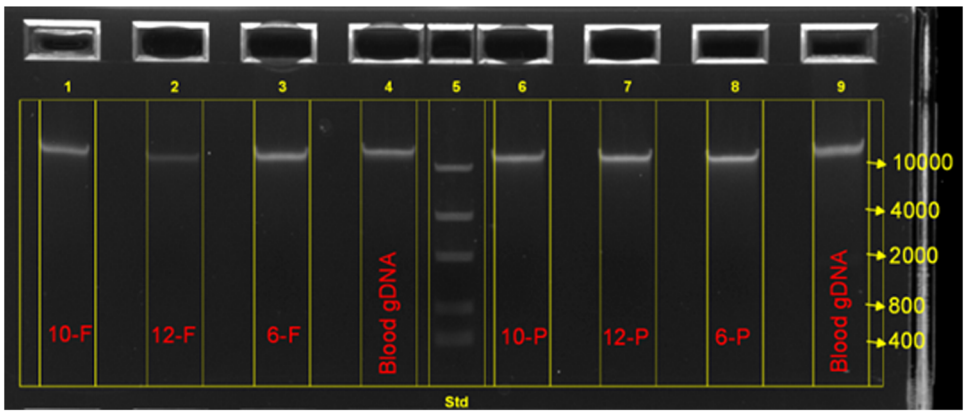

Fig. 1 Electropherogram showing the quality of genomic DNA derived from dried blood spot samples. Image of electrophoresis of genomic DNA from dried blood samples (DBS) 10-FTA, 12-FTA, 6-FTA, 10-PC, 12-PC, 6-PC and blood in a 0.8% agarose gel. Molecular weight standards are shown (nucleotides). A single high-molecular weight band is observed, with no apparent DNA degradation.

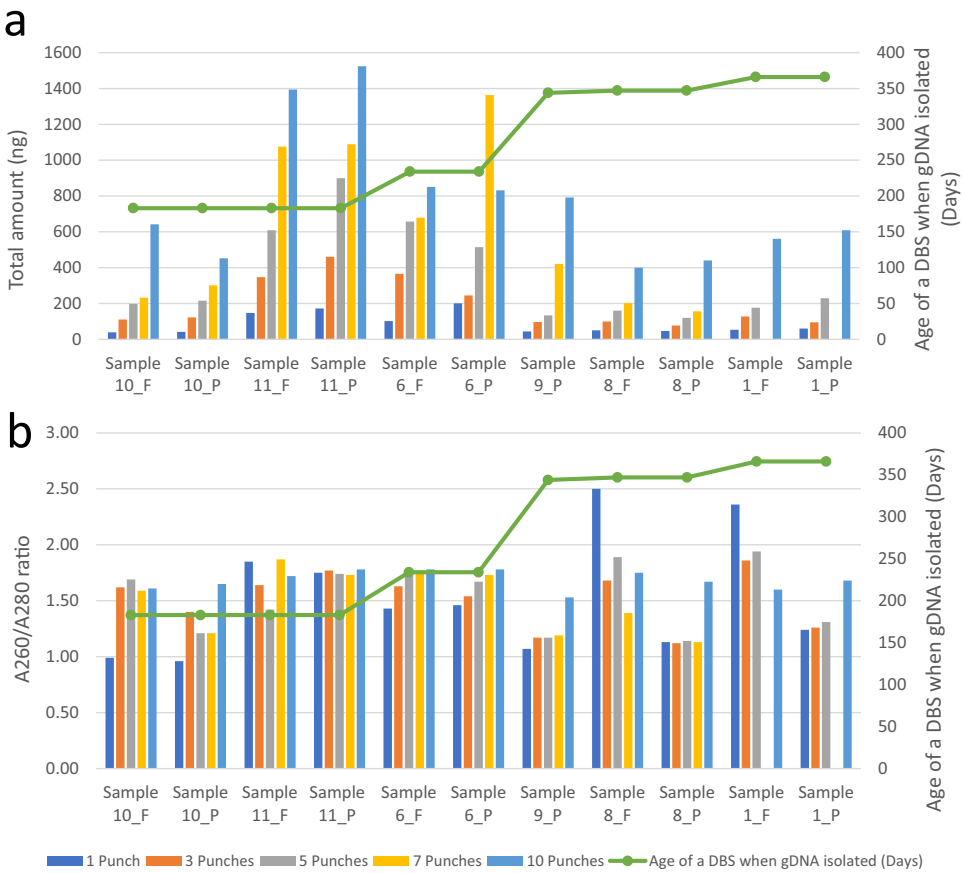

Fig. 2 Relationship between genomic DNA yield, purity, and number of blood spot punches used for DNA extraction. Relationship of number of DBS punches and genomic DNA yield (**a**) and A260/A280 ratio (**b**) for 11 DBS. Samples were 1, 6, 8, 9, 10, and 11. DBS were generated with both FTA and PC filter papers, with the exception of sample 9 (PC only).

gDNA per extraction. The DNA yield decreased with the length of storage of manufactured DBS (Fig. 2a). The A260/A280 ratio did not vary with age of DBS (Fig. 2b).

Quality of sequencing libraries generated from DBS-derived genomic DNA

All 63 gDNA samples yielded sequencing libraries without requirement for PCR (Table 2). There were no library failures. The range of library yield was 3–18 nM, and mean yields were 7.2 and 7.8 nM using Illumina and KAPA methods ($p > 0.05$), respectively. The average library yields from gDNA prepared from EDTA-blood of the same individuals were significantly greater using Illumina and KAPA methods ($p < 0.01$). There was no difference between library yield with FTA or PC DBS with Illumina or KAPA methods ($p > 0.05$; Supplementary Table 2).

Table 2. Quality of diagnostic WGS of 25 blood samples (controls) compared with WGS from 63 DBS made from those blood samples and 29 archived California newborn DBS.

Sample type	Diagnostic WGS samples				Archived CA DBS
	Blood	Blood	DBS	DBS	DBS
Library preparation method	Illumina	KAPA	Illumina	KAPA	Illumina
DBS punches available	n.a.	n.a.	n.a.	n.a.	10.1
WGS performed	24	48	24	39	29
Average raw WGS Yield (Gb)	148	161	139	160	135
Average reads mapped	99.00%	98.60%	97.80%	98.30%	99.30%
Average duplicate Reads	11.80%	10.80%	12.60%	11.00%	9.60%
Mean insert size (bp)	422	383	423	306	421
Average genomic coverage	41	43	37.6	43.2	38.5
Average coverage MIM gene coding domains	41.1	45	36	42.6	36.4
Average MIM genes with ≥10X coverage of 100% coding domain nucleotides	95.40%	96.60%	95.00%	93.60%	91.60%
Average coding variants	25,255	25,078	25,525	25,170	27,788
Average SNVs	3,862,520	3,996,783	3,951,683	3,963,326	3,997,674
Average indels	948,453	931,744	970,186	940,634	987,779
Average mitochondrial genome coverage	10,598	10,820	3,448	8,197	3,112
Average copy number variants overlapping MIM genes	23	22	12	237	9

Two library preparation methods were used (Illumina PCR-free genomic library preparation and KAPA HyperPlus). *MIM* Mendelian Inheritance in Man, *CA* California.

Quality of whole genome sequences derived from DBS

The 63 genomic DNA libraries were sequenced on NovaSeq instruments with S1, S2, or S4 flow cells. Quality metrics of WGS from DBS and EDTA-blood were similar (Supplementary Table 3, Supplementary Table 4, Supplementary Table 5). The average proportion of Q30 nucleotides of DBS libraries on S1, S2, and S4 flow cells were 93.7, 92.9, and 89.5%, respectively, compared with 94.1, 92.9, and 87.8%, respectively, for the corresponding EDTA-blood libraries ($p > 0.05$; Supplementary Table 3). The average nucleotide error rates of DBS libraries on S1, S2, and S4 flow cells were 0.20, 0.18, and 0.22%, respectively, compared with 0.17, 0.18, and 0.27%, respectively, for the corresponding EDTA-blood libraries ($p > 0.05$; Supplementary Table 3).

The uniformity of WGS coverage in DBS was assessed by examining GC bias (defined as relative difference in average coverage of GC-rich regions (62%) to that of regions with the modal human genome GC-content (38%)). The range of GC bias required to pass quality control for clinical libraries made from blood samples was −0.25 to 0.25. 92.5% (49/53) of DBS WGS prepared with the Illumina method and 87.2% (34/39) DBS WGS prepared with the KAPA method were within this range (Supplementary Table 4, Supplementary Table 6). The uniformity of sequence coverage of WGS was also assessed by the standard deviation of coverage normalized to the average coverage and the total length of all sequenced reference genome nucleotides (Supplementary Table 4 and Supplementary Table 7). There were no significant differences seen between fresh blood, manufactured DBS, or archived DBS in these measures ($p > 0.05$). The average standard deviation of coverage normalized to average coverage was 0.19 for WGS from EDTA-blood (range 0.17–0.20, Supplementary Table 4), 0.23 for WGS from manufactured DBS (range 0.19–0.37, Supplementary Table 4), and 0.23 in California Department of Public Health DBS (range 0.19–0.27, Supplementary Table 7). The average mappable genome length was 2.67 GB for all groups.

DNA damage (cytosine deamination causing C > T transitions) was assessed using several metrics: Firstly, an indirect measure was SNV concordance between WGS of fresh blood and DBS from the same individuals (Cohort 1, Supplementary Table 8). It was over 92.6% (Table 3, Supplementary Table 8). Secondly, there were no statistically significant differences in transition to transversion (Ti/Tv) ratio between WGS from DBS archived by the California Department of Public Health DBS (average 2.03 ± 0.01 [standard deviation], Supplementary Table 7), fresh blood (2.03 ± 0.01) or manufactured DBS with Illumina or KAPA methods (2.03 ± 0.01; Supplementary Tables 4 and 7). Thirdly, the rate of cytosine deamination was measured by calculating C>T+G>A/T>C+A>G variant ratios (Supplementary Tables 4 and 7). There were no significant differences between WGS from fresh blood, manufactured DBS, or California Department of Public Health archived DBS.

Quality of alignment and variant calling of whole genome sequences from DBS

Approximately 120 Gb of WGS was generated for 63 DBS-derived libraries and matched blood samples with the KAPA and Illumina library preparation methods. The average genomic coverage of WGS with KAPA was greater than that of Illumina (37.6-fold vs 43.2-fold, $p < 0.01$; Supplementary Table 4). However, the proportion of OMIM genes in which 100% of coding domain nucleotides had coverage ≥10×, an important measure of the ability to call heterozygous variants with confidence, did not differ between these methods (95.0% vs 93.6%, $p > 0.05$; Supplementary Table 4). The average genomic coverage of WGS for the corresponding EDTA-blood libraries was similar to that of the DBS-derived KAPA method (43.0-fold, $p > 0.05$, Supplementary Table 5). However, the proportion of OMIM genes in which 100% of coding domain nucleotides had ≥ 10× coverage was greater in the corresponding EDTA-blood libraries (96.6%, $p < 0.01$; Supplementary Table 5). There were not significant differences between FTA and PC DBS cards in either of these quality metrics using either library preparation method ($p > 0.05$, Supplementary Tables 4 and 5). As expected, the proportion of OMIM genes with ≥10× coverage of the complete coding domain increased with the depth of WGS (Fig. 3). WGS with DBS-derived Illumina libraries and matched blood samples with the KAPA and Illumina methods had very similar distributions of proportions of OMIM genes with ≥10× coverage of the complete coding domain. However, DBS-derived

Table 3. Concordance of single nucleotide and insertion-deletion nucleotide variants (SNVs and indels) in WGS from 25 blood samples and 63 DBS derived from those samples.

Sample Type	Library preparation type	Number of WGS	Average WGS SNVs	WGS SNV concordance: DBS vs blood	Average WGS indels	WGS indel concordance: DBS vs blood
Blood	KAPA	48	3,122,510		510,152	
Blood	Illumina	24	3,109,958		507,375	
FTA DBS	KAPA	12	3,134,167	99.20%	513,429	96.60%
FTA DBS	Illumina	16	3,103,061	99.30%	507,910	96.90%
PC DBS	KAPA	26	3,113,230	99.20%	498,174	96.00%
PC DBS	Illumina	8	3,143,100	99.30%	505,628	97.00%

Libraries were prepared from two types of DBS cards (FTA and Protein Saver [PC]) using two preparation methods (Illumina PCR-free and KAPA HyperPlus).

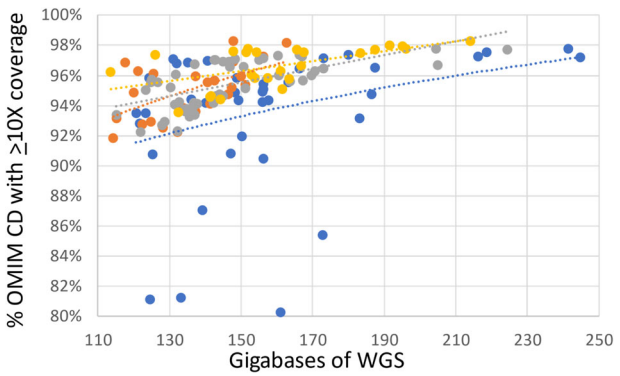

Fig. 3 Relationship between sequence depth and proportion of genes with complete coverage. Relationship between percentage of Mendelian Inheritance in Man (MIM) genes with at least 10-fold coverage of all coding domain nucleotides and amount of WGS. Shown are WGS from DBS prepared with the Illumina method (orange) or KAPA method (sky blue) and matched blood samples prepared with the Illumina method (gray) or KAPA method (yellow). Dotted trend lines are shown.

KAPA libraries had much more variable distributions of proportions of OMIM genes with ≥10× coverage of the complete coding domain. Thus, DBS-derived KAPA libraries required generation of approximately 60 GB more WGS than DBS-derived Illumina libraries or blood-derived KAPA or Illumina libraries to provide ≥10× coverage of the complete coding domain of 92% of OMIM genes. SNV and nucleotide indel variant call accuracy was evaluated by comparing average SNV and indel concordance between WGS from DBS and blood in the same individuals. Mean concordance was 92.6% (range 87.2–94.6%) (Supplementary Table 8). An average of 2.3% of variants were unique to blood samples, and 5.1% variants were unique to DBS (Supplementary Table 8). We examined twenty random discordant calls in WGS from DBS and blood in a proband/sibling/father trio (Supplementary Figs. 2–16). Almost all discrepant variants occurred in regions that were either difficult to sequence, align, or variant call due to repetitive sequences (such as LINE1, Alu, FLAM, MER, or MStB1 endogenous retroviral elements), GC- or AT-rich or homopolymer-containing regions, or had more than one overlapping variant.

Diagnostic performance of DBS-derived WGS

We evaluated the diagnostic recall of WGS performed on DBS in the 19 individuals who had at least one variant reported clinically (Table 1). These were assessed manually in the Integrated Genome Viewer (IGV), as well as by our standard annotation (Enterprise, Fabric Genomics), variant calling, and analysis pipeline (Supplementary Table 1 and Supplementary Figs. 1–16). All 13 SNVs, 2 indels, 9 SV-deletions, and 4 SV-insertions were recapitulated in DBS-derived WGS (Table 1 and Supplementary Figs. 1–16).

WGS from DBS obtained from the California State Biobank

Having established that high quality WGS was possible from manufactured DBS, we sought to examine whether the same methods were effective with 29 randomly selected DBS samples obtained from the California State Biobank. These DBS were collected between 2/22/2000 and 04/06/2020. Six to twelve punches were available per sample (average 10, Supplementary Table 6). Sixteen DBS were FTA, six were PerkinElmer filter paper, and nine DBS were of unknown type (Supplementary Table 1). All DBS yielded high molecular weight DNA (mean 645 ng, range 331–1050 ng, SD 189.74 ng, Supplementary Table 6). The resultant gDNAs yielded sequencing libraries that passed quality control and produced high-quality WGS (Supplementary Table 6). The average sequencing yield was 135 GB (SD 25) (Supplementary Table 6). Resultant average genomic coverage was 38.6-fold (range 27.8–56.6, SD 7.0), while the average coverage of coding domains of OMIM genes was 36.4 (range 25.0–56.7, SD 7.0; Supplementary Table 6). The average percentage of OMIM genes in which 100% of coding domain nucleotides had >10× coverage was 91.6% (Supplementary Table 6). 59% (17/29) of the samples met our standard, clinical quality metric for this measure (>95% of OMIM genes with >10× coverage).

Effect of DBS age and paper type

Statistical analyses were performed on the second cohort of archived DBS WGS to investigate the potential effects of DBS sample age and paper type used. There was no significant correlation between age of the DBS sample and WGS yield, genome coverage, coverage of OMIM genes, or mitochondrial coverage (Supplementary Fig. 1). Nor was there a significant effect of filter paper type on WGS yield or genome coverage. In a previous study, the amount of input DNA had affected the proportion of duplicate reads and the overall coverage of WES[20]. We, therefore, investigated the potential correlation using Pearson's product-moment correlation. We found no significant correlation between the amount of input DNA and proportion of duplicate reads or overall coverage (Supplementary Tables 2 and 6).

DISCUSSION

Archived DBS from national and state screening programs worldwide present an enormous resource for investigating the etiology of pediatric genetic diseases. Here, we demonstrated that high-quality WGS data can be obtained from archived DBS stored for up to 20 years. We also demonstrated in a panel of samples that WGS from DBS identified all of the SNV, indel, SV-deletion,

and SV-duplication findings that had been reported clinically from prior WGS from blood samples, as well as a 92.6% concordance of SNV and small indel calls between WGS from blood and DBS. The methods described herein are simple and use generally available kits. They will enable researchers and clinical laboratories to utilize DBS for PCR-free WGS. DNA isolation from DBS took less than 90 min. Very little degradation of DNA extracted from DBS was observed by agarose gel electrophoresis (Fig. 1). DNA purity (A260/A280 ratio) was slightly lower than that from whole blood but acceptable for WGS, as confirmed by secondary WGS QC metrics and tertiary analysis results. Avoidance of PCR is important for optimal analytic performance of WGS, particular for SVs[28–30]. Compared with gDNA derived from EDTA blood from the same individuals, the library yields of DBS were lower but the primary, secondary, and tertiary analysis of WGS passed quality control criteria. The average coverage of OMIM genes was 36.0× and 42.6× for the Illumina and KAPA preparation methods, respectively, in samples analyzed at RCIGM, and 36.4x in DBS obtained from the California Biobank. This coverage supports confident heterozygous variant calling. There was no observed difference in the quality of WGS associated with sample age or DBS filter paper type. While both the KAPA and Illumina methods worked well, there was much greater variability in the proportion of OMIM genes with at least 10-fold coverage of their coding domain nucleotides with the former. Thus, the Illumina method more consistently yielded adequate disease gene coverage, which is important in clinical production WGS.

The ability to perform high quality PCR-free WGS from archived DBS is significant for several reasons. First, there are over 18 million samples in the California Biobank that encompass a full range of ethnic, socioeconomic, regional, and temporal diversity within the state. This represents an enormous resource for genomic studies. Large-scale sequencing studies of these DBS have the potential to reveal new genotype-phenotype associations and further our knowledge of genetic diseases in human populations. Currently, consent for sample storage is obtained at the time of collection, and these DBS samples are then stored indefinitely in the California State Biobank. This wealth of stored biospecimen data may support large-scale studies to investigate the epidemiology of rare genetic disease in a way that has not previously been feasible. For example, the San Diego Study of Outcomes for Mothers and Infants (SOMI) dataset currently supports linkage of vital statistics, DBS samples, death certificates, and hospital records for infants born in San Diego County after 2007. These types of datasets will enable the linkage of epidemiological data with individual-level molecular diagnoses. Second, there is growing interest in the potential of WGS for expanded newborn screening[12,20,31–35]. Advances in WGS technology have recently made fully automated, diagnostic WGS and management guidance possible in 13.5 h[34]. Those methods are compatible with the WGS from DBS described herein[33]. Thus, autonomous, expanded newborn screening by WGS of DBS is conceivable[32,33]. In screening mode, WGS interpretation would be limited to known pathogenic and likely pathogenic variants, which could be detected without the need for trio sequencing—confirmatory testing could incorporate parental samples if warranted, as is done currently[32,33]. Adding new conditions to the recommended uniform screening panel is costly, and each additional assay requires independent state or federal legislative approval[35]. In California, 84 conditions are currently screened and 1 in 600 newborns has a positive NBS result[11]. In contrast, WGS has the potential to screen for known pathogenic and likely pathogenic variants in over 7200 disorders at once, including those that can cause sudden infant death[14,15,32–34]. There are many rare genetic diseases, such as pyridoxine-dependent epilepsy, which may have effective treatments and meet the Wilson and Junger criteria for inclusion in newborn screening[33,36–39]. Newborn screening by WGS has potential to reduce morbidity and mortality associated with these conditions. Certain genetic diseases, such as spinal muscular atrophy, also rely on the particular variants identified to guide specific management. As shown herein in two cases, WGS provides this level of analytic performance. Here, we have demonstrated that PCR-free WGS is feasible for use on archived DBS collected from NBS, expanding the potential of their use for investigating the prevalence of rare genetic diseases.

METHODS

Study design

This study received a waiver of consent from the Rady Children's Hospital and University of California – San Diego (UCSD) institutional review boards (IRB) and was undertaken as a quality improvement project. DBS samples were from two cohorts: The first consisted of 25 children aged less than 18 years, of male and female sex, who received WGS from whole blood at the Rady Children's Institute for Genomic Medicine (RCIGM) between January 2018 and June 2019 for diagnosis of a suspected genetic disease. WGS was performed in laboratories accredited by the College of American Pathologists (CAP) and certified through Clinical Laboratory Improvement Amendments (CLIA). They had either been consented under various research protocols approved by the UCSD IRB and the Johns Hopkins IRB or were sample retains from clinical diagnostic testing. Blood spots were created using two paper types (filter chemical treated and non-treated) and stored for these individuals. Two cases had additional family members available for analysis: sample 6 had an affected sibling and father available, and sample 8 had father's genome available (Table 1). WGS data from DBS were compared with prior, clinical WGS data from whole blood. The second cohort were randomly selected, de-identified California Department of Public Health (CDPH) DBS samples from newborns of male and female sex without associated clinical data. No IRB approval was required from the CDPH IRB for use of de-identified DBS to determine feasibility of methods.

Dried blood spots

The first cohort consisted of DBS generated with 40 µl EDTA-whole blood from 25 individuals. Twenty nine DBS sets (each set containing multiple cycles or spots, see Supplementary Table 1 for details) were prepared with Whatman NUCLEIC-CARD™ matrix (FTA) (ThermoFisher, Catalog #: 4473975) and 34 DBS sets were prepared with 903 Protein Saver 903 Cards (PC) (GE Healthcare, Catalog #:10534612). Preliminary data showed similar final yield of gDNA from spotting 80 and 40 µl of blood per disk. Multiple spots were made per individual (Supplementary Table 1). The DBS cards were kept at room temperature for at least 2 h before they dried. Dried cards were stored inside a desiccator at room temperature for later use.

The second cohort consisted of 29 randomly selected, de-identified, anonymized DBS obtained from the California Biobank Program. Two DBS were collected for each blood sample, and a variable number of punches was available (Supplementary Table 6). All samples had >7 punches available, and 6 punches were used for each sample to maintain consistency (Supplementary Tables 1 and S6). Samples were archived after collection and stored with desiccant at −20 °C. They represented a range of filter types and years in storage.

Genomic DNA isolation

Each DBS disc was visually examined prior to isolation to ensure no damage had occurred and that there was full absorption on the paper. For cohort 1, two different lysis protocols were used for the four tests performed. Lysis protocol 1 was performed as

following: six 3 mm^2 punches from a DBS specimen were manually collected in a 1.5 mL Eppendorf tube and mixed with 2 µl proteinase K, 30 µl lysis buffer (DNA Flex Lysis Reagent Kit, catalog # 2018706 or QIAgen catalog #19075), and 268 µl nuclease free water. The sample tube then was incubated at 56 °C for 15 min in a thermomixer set at 1000 rpm. Lysis protocol 2 was performed as following: ten 3 mm^2 punches were mixed with 4 µl proteinase K, 40 µl lysis buffer and 356 µl nuclease free water. The sample tube was then incubated at 56 °C for 60 min in a thermomixer set at 2000 rpm. The titration experiment applied different amounts of input at 1, 3, 5, 7, and 10 punches per extraction. For cohort 2, lysis protocol 1 was used for all samples.

For both cohorts, after incubation, the punches/reagent mixture was briefly spun down and the supernatant was carefully transferred into a new Eppendorf tube without disturbing the DBS. 135 µl (for Lysis Protocol 1) or 175 µl (for Lysis Protocol 2) of well-mixed, room temperature normalized KAPA pure beads (Roche/KAPA Biosystems, Catalog #: KK8002) was added into the tube and the solution was mixed by rotating the tube on a rotator (or equivalent) for 15 min at room temperature. The sample tube then was placed on a magnet bar (or equivalent) for 5 min, the supernatant was discarded, and the pellet was washed twice with 500 µl 80% ethanol. The sample/pure beads were air-dried for a few minutes at room temperature before genomic DNA was eluted using 20–40 µl elution buffer (10mM Tris-HCl, pH 8 to 8.5). Genomic DNA (gDNA) then was quantified and qualified using Picogreen assay and Nanodrop A260/A280 assays (ThermoFisher), by following manufacturer's protocols[40,41]. Electrophoresis using 0.8% E-gel (ThermoFisher, catalog # A25798) was performed for a subset of selected gDNA samples to evaluate the integrity of the extracted gDNA. The manual isolation of the lysis 1 and lysis 2 protocol took approximately about 60 and 100 min respectively. The age of DBS from the time it was made to the time when the gDNA was isolated ranged from 1 day to 366 days. Average age of DBS at time of gDNA extraction was 117 days. The integrity of extracted gDNA was examined by electrophoresis using 0.8% agarose gels (ThermoFisher).

Genome sequencing library preparation

PCR-free libraries were prepared with either DNA PCR-free (Tagmentation) Prep kits (Illumina) or KAPA HyperPlus PCR-free library kits (Roche, abbreviated KAPA herein) for the 63 manufactured DBS as detailed below (Supplementary Table 1)[42,43]. Both the Illumina and KAPA method for PCR-free library took approximately 3 h. For the archived DBS from California Department of Public Health (CDPH), WGS libraries were prepared using Illumina prep kits.

For Illumina PCR-free Tagmentation (Cat#: 20041855)[42], an average of 286 ng gDNA in 10 mM tris-HCl (pH 8 or 8.5) solution was isolated from each DBS and incubated with 10 µl tagmentation buffer and 10 µl bead-linked transposomes at 41 °C for 5 min. 10 µl stop buffer was added and well mixed, then incubated at room temperature for 5 min. The sample mixture was placed on a magnet bar or plate until the solution was clear, then about 60 µl supernatant was discarded and 150 µl wash buffer was added while the sample was kept on the magnet bar or plate. The 150 µl wash buffer was then removed. For the ligation step, 45 µl extension ligation mix and 5 µl index adaptor were both added, and the sample mixture was incubated at 37 °C for 5 min and 50 °C for 5 min. Using 75 µl wash buffer, the products were washed while keeping the sample mixture on the magnet bar or plate. 75 µl wash buffer was then discarded and 47 µl sodium hydroxide (2N) was added into the sample and incubated at room temperature for 5 min. The sample mixture was placed on the magnet bar or plate again and the supernatant was removed. Finally, beads-based double size selection was performed to ensure the fragment size of each sample was within from 450 to 650 bp, following manufacturer protocol.

Libraries were constructed using the Roche KAPA HyperPlus kits[43] (Cat#: KK8515) with an average 400 ng of extracted gDNA. For enzymatic fragmentation, the gDNA was normalized to 30 µl in the suspension buffer. 20 µl fragmentation mixture containing 5 µl diluted conditioning buffer (13.5 µl original conditioning buffer in 86.5 µl nuclease free water), 5 µl fragmentation buffer and 10 µl fragment enzyme were added to each sample well and the plate was incubated at 37 °C for 8 min. After incubation, a pre-made 10 µl End-Repair (ER) and A-tailing (AT) mixture (7 µl End-Repair/A-tailing buffer, 3 µl enzyme) was immediately added into each sample well and the plate was incubated at 65 °C for 30 min. In the ligation step, 48 µl ligation master mix (30 µl ligation buffer, 10 µl DNA ligase, 8 µl PCR-grade water) and 2 µl dual index adapter (IDT, San Diego, CA) were added to each well containing sample/ER/AT mix, and the plate was incubated at 20 °C for 30 min in a thermocycler. Samples were cleaned up using 1× SPRI (Solid Paramagnet Reversible Immobility) beads and 80% ethanol. Finally, beads-based double size selection was performed to ensure the fragment size of each sample was within from 450 to 650 bp, following the manufacturer's protocol. The fragment size of a DNA library sample was measured using Agilent DNA High Sensitivity NGS Fragment Analysis Kit (Agilent, Catalog#: DNF-474-0500) to ensure it was between 300 and 600 bp.

Whole genome sequencing

The concentration of ligated fragments was quantified with KAPA Library Quantification Kits for Illumina platforms (Roche/KAPA Biosystems, Catalog#: KK4824) on Roche LightCycler 480 Instrument (Roche, Basel, Switzerland). Libraries with concentration >3 nM and acceptable fragment size passed quality control and were sequenced on Illumina Novaseq 6000 instruments. Libraries from the 63 manufactured DBS were pooled at equal molarity for a final loading concentration between 400 and 450 pM as follows: S1 flow cell, 2.5 libraries, S2 flow cell, 5–6 libraries, or S4 flow cell, 24 libraries. The pooled libraries were denatured with 0.2 N sodium hydroxyl for 8 min, followed by addition of 400 mM Tris-HCl (pH 8.5). The flow cells were loaded on the Illumina Novaseq 6000 with a read length of 2×101 or 2×151 (IDT dual indexing, Cat#:263582653) for cohort 1, or 2×250 or 2×100 for cohort 2. Quality metrics for passing WGS were Q30 ≥ 80%, error rate ≤3% and >120 Gb.

Secondary analysis of whole genome sequences

WGS from the manufactured DBS were aligned to human genome assembly GRCh37 (hg19) and variants identified with the Illumina DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT Platform (Illumina, Supplementary Table 1). WGS from archived DBS were aligned to human genome assembly GrCh37 (hg19) and nucleotide variants were identified with the DRAGEN platform (Illumina, Supplementary Table 1). Samples were run using different versions of DRAGEN as described in Supplementary Table 1. For each update of DRAGEN, a verification process was applied to ensure the quality of DRAGEN variant calling as it pertains to the RCIGM clinical diagnostic standards. Briefly, for each DRAGEN upgrade, VCF concordance between old and new DRAGEN results are verified to have >99% F2-measure as calculated by vcfeval[44]. Furthermore, quality control metrics, such as mapping rate and average genomic coverage are verified to have >98% concordance between old and new DRAGEN results. Structural variants were identified with Manta and CNVnator (using DNAnexus). Structural variants were filtered to retain those affecting coding regions of known disease genes and with allele frequencies <2% in the RCIGM database. All samples underwent a battery of quality controls, including: (1) sample identity tracking (STR/CODIS) from orthogonal inputs (capillary electrophoresis using

Genetic Analyzer ThermoFisher 3500xl) and in silico STR from WGS; (2) <15% duplicate rate, (3) >98% aligned reads rate; (4) Ti/Tv in appropriate range (2.0–2.2); (5) Hom/Het in appropriate range (0.50–0.61); (6) >90% of OMIM genes with >10-fold coverage of every coding nucleotide; (7) sex match; (8) additional technical controls (insert size and others). Coverage uniformity was assessed using the GC bias measure as well as two additional measures: the standard deviation of coverage normalized to average coverage and the total length of all reference genome regions with read coverage. Both of these measures were computed by binning the complete genome coverage into bins of 200 bases at a time (Supplementary Tables 4 and S7). For a small fraction of the sample, these metrics were not retained, and these have been marked as ND (Supplementary Tables 4 and S7).

Surveillance of cross-sample contamination during DBS WGS process

Intra-sample and intra-batch cross contamination were monitored with following measurements: 1. Ensure cleanness of each DBS sample when using a Harris Uni-Core punch to remove a sample disc to a sample tube by punching on a clean paper before working on the next DBS sample according to manufacturer's recommendation[44]. 2. A negative control (NTC) was included during WGS library construction and QC quantification, samples in a batch with NTC contamination would be failed and would not be passed for sequencing. 3. In silico analysis post sequencing, cross-sample contamination was computed and had to pass defined criteria for downstream variant calling.

Concordance analysis between EDTA blood and DBS WGS

Small nucleotide variants concordance analysis for datasets derived from EDTA blood samples and DBS samples were performed according to best practices set forth by Global Alliance for Genomics and Health (GA4GH) Benchmarking Team[45]. Briefly, after generation, the VCF files were compared using the vcfeval software[45]. The GA4GH Benchmarking Team developed standardized performance metrics for genomic variant calls as well as sophisticated variant comparison tools to robustly compare different representations of the same variant, and a set of standard browser extensible data (BED) files describing difficult genome contexts to stratify performance. The GA4GH Benchmarking application requires a truth VCF file (for this study, the EDTA blood sample), the truth confident regions (the GIAB high-confidence BED file for HG002 was used), and the query VCF file (the DBS sample in this study). The GA4GH application returns the count of false negatives (FNs), false positives (FPs), and true positives (TPs) in both standardized VCF and comma-separated value formats. Performance metrics follow the GA4GH standardized definitions, in which genotyping errors are counted both as FP and FN. Precision (also known as positive prediction value (PPV)) was calculated using the following formula: $PPV = TP/(TP + FP)$. Sensitivity was calculated using the following formula: $sensitivity = TP/(TP + FN)$.

Diagnostic utility of WGS from DBS

Nineteen of the 25 individuals in whom DBS were manufactured had received diagnostic results from clinical WGS from whole blood at RCIGM according to American College of Medical Genetics and Genomics (ACMG)/Association of Molecular Pathology (AMP) guidelines. Aligned sequences from DBS-based WGS were viewed in the Integrated Genomics Viewer (IGV) and Fabric Enterprise (Fabric Genomics) to determine whether the variants and diplotypes were recapitulated.

Statistical analysis

A student's t-test was used to compare the means of test and control groups. P values <0.05 were considered statistically significant. Pearson product moment correlations were used to analyze relationships between age of the bloodspot and quality metrics. Two-tailed Fisher's exact tests were performed to compare the effects of different paper types. All analyses were conducted in R v.4.0.3, and visualization was done using the packages ggpubr and ggplot2[46,47].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Whole genome sequencing data are subject to conditions of the IRB protocols and CDPH policies under which the data was generated, and therefore the raw sequencing data is unavailable. For additional summary or aggregate level data, please contact Dr. Stephen Kingsmore, skingsmore@rchsd.org.

CODE AVAILABILITY

DRAGEN v.4.1 is available at https://www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html. Manta v.1.6.0 is available at https://github.com/Illumina/manta. CNVnator v.0.4.1 is available at https://github.com/abyzovlab/CNVnator. vcfeval v.3.12.1 is available at https://github.com/Illumina/hap.py/blob/master/src/python/Haplo/vcfeval.py. IGV v.2.15.4 is available at https://software.broadinstitute.org/software/igv/download. Fabric Enterprise v.6.15.9 is available from Fabric Genomics Inc.

Received: 4 March 2021; Accepted: 5 January 2023;
Published online: 14 February 2023

REFERENCES

1. Berg, J. S. et al. Newborn sequencing in genomic medicine and public health. *Pediatrics* **139**, e20162252 (2017).
2. Therrell, B. L. & Adams, J. Newborn screening in North America. *J. Inherit. Metab. Dis.* **30**, 447–465 (2007).
3. Guthrie, R. & Susi, A. A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants. *Pediatrics* **32**, 338–343 (1963).
4. Pitt, J. J. Newborn screening. *Clin. Biochem. Rev.* **31**, 57–68 (2010).
5. NewSTEPS Annual Report 2020. https://www.aphl.org/aboutAPHL/publications/Documents/NewSTEPs_Annual_Report_9_22_21.pdf (2021).
6. Baker, M. W. et al. Improving newborn screening for cystic fibrosis using next-generation sequencing technology: a technical feasibility study. *Genet. Med.* **18**, 231–238 (2016).
7. Pollitt, R. J. Newborn blood spot screening: new opportunities, old problems. *J. Inherit. Metab. Dis.* **32**, 395–399 (2009).
8. Newborn screening: a blueprint for the future executive summary: newborn screening task force report. *Pediatrics* **106**, 386 (2000).
9. California Newborn Screening Program Open Data, Screening Disorders by California Region. https://data.chhs.ca.gov/dataset/b4810e9c-c364-4231-aef8-66ee14d9213e/resource/fbb8fdcd-38d5-4068-9e8a-9b76f22bb006/download/newborn-screening-disorders-californiaregion-2009-2019_final.csv (2020).
10. Hollegaard, M. V., Grove, J., Thorsen, P., Nørgaard-Pedersen, B. & Hougaard, D. M. High-throughput genotyping on archived dried blood spot samples. *Genet. Test. Mol. Biomark.* **13**, 173–179 (2009).
11. California Newborn Screening Program Open Data, Recommended Uniform Screening Panel (RUSP) Core Conditions. https://data.chhs.ca.gov/dataset/b4810e9c-c364-4231-aef8-66ee14d9213e/resource/6b43242e-e12e-409e-951f-5dc5cebf43db/download/rusp-uniform-screening-panel_july2018.pdf (2020).
12. Adhikari, A. N. et al. The role of exome sequencing in newborn screening for inborn errors of metabolism. *Nat. Med.* **26**, 1392–1397 (2020).
13. Rothwell, E., Johnson, E., Riches, N. & Botkin, J. R. Secondary research uses of residual newborn screening dried bloodspots: a scoping review. *Genet. Med.* **21**, 1469–1475 (2019).

14. Kingsmore, S. F. et al. Mortality in a neonate with molybdenum cofactor deficiency illustrates the need for a comprehensive rapid precision medicine system. *Cold Spring Harb. Mol. Case Stud.* **6**, a004705 (2020).
15. Sanford, E. et al. Postmortem diagnosis of PPA2-associated sudden cardiac death from dried blood spot in a neonate presenting with vocal cord paralysis. *Cold Spring Harb. Mol. Case Stud.* **6**, a005611 (2020).
16. Kingsmore, S. F. et al. Measurement of genetic diseases as a cause of mortality in infants receiving whole genome sequencing. *NPJ Genom. Med.* **5**, 49 (2020).
17. Hollegaard, M. V. et al. Archived neonatal dried blood spot samples can be used for accurate whole genome and exome-targeted next-generation sequencing. *Mol. Genet. Metab.* **110**, 65–72 (2013).
18. Poulsen, J. B. et al. High-quality exome sequencing of whole-genome amplified neonatal dried blood spot DNA. *PLoS One* **11**, e0153253 (2016).
19. Hollegaard, M. V. et al. Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source. *BMC Genet.* **12**, 58 (2011).
20. Bassaganyas, L. et al. Whole exome and whole genome sequencing with dried blood spot DNA without whole genome amplification. *Hum. Mutat.* **39**, 167–171 (2018).
21. Clark, M. M. et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci. Transl. Med.* **11**, eaat6177 (2019).
22. Farnaes, L. et al. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom. Med.* **3**, 10 (2018).
23. Miller, N. A. et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med.* **7**, 100 (2015).
24. Kingsmore, S. F. et al. A randomized, controlled trial of the analytic and diagnostic performance of singleton and trio, rapid genome and exome sequencing in Ill Infants. *Am. J. Hum. Genet.* **105**, 719–733 (2019).
25. Winkel, B. G. et al. Whole-genome amplified DNA from stored dried blood spots is reliable in high resolution melting curve and sequencing analysis. *BMC Med. Genet.* **12**, 22 (2011).
26. Hollegaard, M. V. et al. Whole genome amplification and genetic analysis after extraction of proteins from dried blood spots. *Clin. Chem.* **53**, 1161–1162 (2007).
27. Sok, P. et al. Utilization of archived neonatal dried blood spots for genome-wide genotyping. *PLoS One* **15**, e0229352 (2020).
28. Kozarewa, I. et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
29. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
30. Sato, M. P. et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Res* **26**, 391–398 (2019).
31. Trier, C. et al. Next-generation sequencing of newborn screening genes: the accuracy of short-read mapping. *NPJ Genom. Med.* **5**, 36 (2020).
32. Kingsmore, S. F. BeginNGS Consortium. Dispatches from Biotech beginning BeginNGS: Rapid newborn genome sequencing to end the diagnostic and therapeutic odyssey. *Am. J. Med. Genet. C Semin. Med. Genet.* **190**, 243–256 (2022).
33. Kingsmore, S. F. et al. A genome sequencing system for universal newborn screening, diagnosis, and precision medicine for severe genetic diseases. *Am. J. Hum. Genet.* **109**, 1605–1619 (2022).
34. Owen, M. J. et al. An automated 13.5 hour system for scalable diagnosis and acute management guidance for genetic diseases. *Nat. Commun.* **13**, 4057 (2022).
35. Bronstein, M. G., Pan, R. J., Dant, M. & Lubin, B. Leveraging evidence-based public policy and advocacy to advance newborn screening in California. *Pediatrics* **143**, e20181886 (2019).
36. Phoon, C. K. L. et al. Sudden unexpected death in asymptomatic infants due to PPA2 variants. *Mol. Genet. Genom. Med.* **8**, e1008 (2020).
37. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A.* **179**, 885–892 (2019).
38. Wilson, J. M. W. & Jungner, G. *Principles and Practice of Screening for Disease* (World Health Organization, 1968).
39. Kingsmore, S. F. BeginNGS Consortium. Dispatches from Biotech beginning BeginNGS: Rapid newborn genome sequencing to end the diagnostic and therapeutic odyssey. *Am. J. Med. Genet. C* **190**, 243–256 (2022).
40. Thermo Scientific. Picogreen Assay for dsDNA. http://tools.thermofisher.com/content/sfs/manuals/PicoGreen-dsDNA-protocol.pdf. Accessed 2/22/2021. (2008).
41. Thermo Scientific. NanoDrop 2000/2000c Spectrophotometer, V1.0 User Manual. https://www.thermofisher.com/order/catalog/product/ND-2000#/ND-2000. Accessed 2/22/2021. (2009).
42. Illumina, D. N. A. *PCR-Free Prep, Tagmentation.* https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/dna-pcr-free-prep.html. Accessed 2/6/2022. (2022).
43. KAPA Biosystems. KAPA HyperPlus Kit. https://www.n-genetics.com/products/1104/1023/17277.pdf. Accessed 2/22/2021. (2020).
44. Cleary, J. G. et al. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **21**, 405–419 (2014).
45. The Global Alliance for Genomics and Health Benchmarking Team et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
46. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
47. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer International Publishing, 2016).

ACKNOWLEDGEMENTS

This work was supported by grant HD101540 from NICHD.

AUTHOR CONTRIBUTIONS

Y.D., M.O., S.N., and S.K. designed the study. Y.D., S.B., Z.B.O., L.V.D.K., J.L., and Z.Z. performed sequencing and analysis. Y.D., S.K., and M.O. wrote the manuscript and made the figures and tables. Y.H.K., N.V., S.B., K.H., and K.C. performed bioinformatic analyses. S.B. and M.B. aided in study design and data analysis. R.B., G.B., and C.C. obtained DBS from the California State Biobank and provided logistical and analytic support. All authors reviewed the final version. S.K. and C.C. acted as the principal investigator on the study, edited the manuscript, and contributed to study design along with J.G.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-023-00349-w.

Correspondence and requests for materials should be addressed to Mallory Owen or Stephen F. Kingsmore.

Reprints and permission information is available at http://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 Author(s) 2023