Introduction

Diamond-Blackfan anemia (DBA [MIM: 105650]), originally termed congenital hypoplastic anemia, is an inherited bone marrow failure syndrome estimated to occur in 1 out of 100,000 to 200,000 live births. A consensus clinical diagnosis for DBA suggests that individuals with this disorder should present within the first year of life and have normochromic macrocytic anemia, limited cytopenias of other lineages, reticulocytopenia, and a visible paucity of erythroid precursor cells in the bone marrow. Nonetheless, an increasing number of cases that fall outside of these strict clinical criteria are being recognized. Treatment with corticosteroids can improve the anemia in 80% of case subjects, but individuals often become intolerant to long-term corticosteroid therapy and turn to regular red blood cell transfusions, the only available standard therapy for the anemia. Currently, a hematopoietic stem cell transplant (HSCT) is the sole curative option, but this procedure carries significant morbidity and is generally restricted to those with a matched related donor. Ultimately, 40% of case subjects remain dependent upon corticosteroids which increase the risk of heart disease, osteoporosis, and severe infections, while another 40% become dependent upon red cell transfusions, the only available standard therapy for the anemia. Currently, a hematopoietic stem cell transplant (HSCT) is the sole curative option, but this procedure carries significant morbidity and is generally restricted to those with a matched related donor. Ultimately, 40% of case subjects remain dependent upon corticosteroids which increase the risk of heart disease, osteoporosis, and severe infections, while another 40% become dependent upon red cell transfusions, the only available standard therapy for the anemia. Currently, a hematopoietic stem cell transplant (HSCT) is the sole curative option, but this procedure carries significant morbidity and is generally restricted to those with a matched related donor.

Diamond-Blackfan anemia (DBA) is a rare bone marrow failure disorder that affects 7 out of 1,000,000 live births and has been associated with mutations in components of the ribosome. In order to characterize the genetic landscape of this heterogeneous disorder, we recruited a cohort of 472 individuals with a clinical diagnosis of DBA and performed whole-exome sequencing (WES). We identified relevant rare and predicted damaging mutations for 78% of individuals. The majority of mutations were singletons, absent from population databases, predicted to cause loss of function, and located in 1 of 19 previously reported ribosomal protein (RP)-encoding genes. Using exon coverage estimates, we identified and validated 31 deletions in RP genes. We also observed an enrichment for extended splice site mutations and validated their diverse effects using RNA sequencing in cell lines obtained from individuals with DBA. Leveraging the size of our cohort, we observed robust genotype-phenotype associations with congenital abnormalities and treatment outcomes. We further identified rare mutations in seven previously unreported RP genes that may cause DBA, as well as several distinct disorders that appear to phenocopy DBA, including nine individuals with biallelic CECR1 mutations that result in deficiency of ADA2. However, no new genes were identified at exome-wide significance, suggesting that there are no unidentified genes containing mutations readily identified by WES that explain >5% of DBA-affected case subjects. Overall, this report should inform not only clinical practice for DBA-affected individuals, but also the design and analysis of rare variant studies for heterogeneous Mendelian disorders.

1Division of Hematology/Oncology, The Manton Center for Orphan Disease Research, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA; 2Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; 3Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA; 4Division of Genetics and Genomics, The Manton Center for Orphan Disease Research, Boston Children’s Hospital, Harvard Medical School, Boston, MA 02115, USA; 5Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; 6Department of Pediatrics, University of Massachusetts Medical School, Worcester, MA 01605, USA; 7Department of Hematology/Oncology, Medical University of Warsaw, Warsaw, Poland; 8Feinstein Institute for Medical Research, Manhasset, NY; 9Division of Hematology/Oncology and Stem Cell Transplantation, Cohen Children's Medical Center, New Hyde Park, NY; 10Hostra Northwell School of Medicine, Hempstead, NY 11030, USA; 11Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 02114, USA; 12Laboratory of Eukaryotic Molecular Biology, Center for Integrative Biology (CBI), University of Toulouse, CNRS, Toulouse, France; 13Murdoch Children’s Research Institute and Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia; 14Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; 15University Paris VII Denis DIDEROT, Faculté de Médecine Xavier Bichat, 75019 Paris, France; 16Laboratory of Excellence for Red Cell, LABEX Gr-Ex, 75015 Paris, France; 17DNA Therapeutics Institute, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605, USA; 18Department of Genetics and Genomic Sciences and The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 19Harvard Stem Cell Institute, Cambridge, MA 02138, USA

*These authors contributed equally to this work
*Correspondence: sankaran@broadinstitute.org (V.G.S.), hanna.gazda@childrens.harvard.edu (H.T.G.)
transfusions which requires regular chelation to prevent iron overload and increases the risk of alloimmunization and transfusion reactions, both of which can be severe co-morbidities.\textsuperscript{2,5}

In contrast to many other rare, presumed monogenic or Mendelian disorders,\textsuperscript{7-9} putative causal genetic lesions have now been identified in an estimated 50\%–60\% of DBA-affected case subjects.\textsuperscript{7} In 1999, mutations in ribosomal protein S19 (RPS19), one of the proteins in the 40S small ribosomal subunit, were identified as the first causal genetic lesions for DBA that explained \~\textasciitilde 25\% of case subjects.\textsuperscript{10} Through the use of targeted Sanger sequencing, whole-exome sequencing (WES), and copy number variant (CNV) assays, putatively causal haploinsufficient variants have been identified in 19 of the 79 ribosomal protein (RP) genes (\textit{RPS}19 [MIM: 603474, 105650], \textit{RPL}5 [MIM: 603364, 612561], \textit{RPS}26 [MIM: 603701, 613309], \textit{RPL}11 [MIM: 604175, 612562], \textit{RPL}35A [MIM: 180468, 612528], \textit{RPS}10 [MIM: 603362, 613308], \textit{RPS}24 [MIM: 602412, 610629], \textit{RPS}17 [MIM: 180472, 612527], \textit{RPS}7 [MIM: 603658, 612563], \textit{RPL}26 [MIM: 603704, 614900], \textit{RPL}15 [MIM: 604174, 615550], \textit{RPS}29 [MIM: 603633, 615909], \textit{RPS}28 [MIM: 603685, 606164], \textit{RPL}31 [MIM: 617415], \textit{RPS}27 [MIM: 603702, 617409], \textit{RPL}27 [MIM: 607526, 617408], \textit{RPL}35, \textit{RPL}18 [MIM: 604179], \textit{RPS}15A [MIM: 603674]), making DBA one of the best genetically defined congenital disorders.

In 2012, through the use of unbiased WES, mutations in \textit{GATA}1 (MIM: 305371, 300835), a hematopoietic master transcription factor that is both necessary for proper erythropoiesis and sufficient to reprogram alternative hematopoietic lineages to an erythroid fate, were identified as the first non-RP mutations in DBA.\textsuperscript{11,12} Further studies on \textit{GATA}1 and other novel genes mutated in DBA, including the \textit{RPS}26 chaperone protein TSR2 (MIM: 300945, 300946),\textsuperscript{13,14} have provided new insights into the pathogenesis of this disorder, suggesting that DBA results from impaired translation of key erythroid transcripts, such as the mRNA encoding \textit{GATA}1, in early hematopoietic progenitors which ultimately impairs erythroid lineage commitment.\textsuperscript{14,18} (This set of 19 RP genes, \textit{GATA}1, and \textit{TSR}2 are henceforth referred to as DBA-associated genes for simplicity, although it is mutations within these genes and not the genes themselves that ultimately cause DBA.)

Given the success of unbiased WES in identifying pathogenic mutations in many Mendelian disorders,\textsuperscript{7,8,11,13,19,20} we recruited and performed sequencing on a large cohort of 472 affected individuals, the size of which is equivalent to 6 years of spontaneous DBA births in the USA, Canada, and Europe, containing individuals with a clinical diagnosis or strong suspicion of DBA. In this report, we describe the results of an exhaustive genetic analysis of this cohort and discuss our experience of attempting to achieve comprehensive molecular diagnoses while limiting false positive reports.

**Material and Methods**

**Diamond-Blackfan Anemia Cohort**

From 1998 until 2018, we recruited a cohort of 472 affected individuals with a clinical diagnosis or strong suspicion of DBA (Table 1). Briefly, 112 affected individuals and their families were recruited through the DBA registry of North America; 73 affected individuals and their families were recruited through the French DBA registry; and 287 affected individuals and their families were recruited from hematological centers and clinics from the USA (176), Poland (67), Turkey (16), and 13 other countries (28) (Table S1). The diagnosis of DBA was based on normochromic, often macrocytic anemia, reticulocytopenia, bone marrow erythroid hyperplasia, and in some individuals, physical abnormalities and elevated erythrocyte adenosine deaminase activity. However, we note that, given the international nature of this cohort, this was not uniformly assessed by any single clinician or center.

The study was approved by the Institutional Review Board at Boston Children’s Hospital. Informed consent was obtained from affected individuals and their family members participating in the study. According to our study protocol, incidental findings that were unrelated to clinical features at presentation were not reported. DNA from whole blood samples and individual derived lymphoblastoid cell lines was obtained for 90\% (427/472) and 10\% (45/472) of individuals, respectively.

**Whole-Exome Sequencing**

By 2010, prior to the widespread adoption of WES, approximately 200 individual DNA samples from the cohort had been screened for mutations in 8 RP genes exclusively by Sanger sequencing. Starting in 2011, all previously collected and newly collected DNA samples underwent both WES and Sanger sequencing: from 2010 to 2017, 11 RP genes and \textit{GATA}1 were screened, and since 2017 16 RP genes and \textit{GATA}1 were screened. A total of 445 affected individuals and 72 unaffected family members underwent whole-exome sequencing at the Broad Institute (dbGAP accession phs000474.v3.p2). Generally, whole-exome sequencing and variant calling was performed as previously reported with several modifications.\textsuperscript{11} Library construction was performed as described in Fisher et al.,\textsuperscript{21} with the following modifications: initial genomic DNA input into shearing was reduced from 3 \( \mu \)g to 10–100 ng in 50 \( \mu \)L of solution. For adaptor ligation, Illumina paired end adapters were replaced with palindrome forked adapters, purchased from Integrated DNA Technologies, with unique 8 base molecular barcode sequences included in the adaptor sequence to facilitate downstream pooling. With the exception of the palindrome forked adapters, the reagents used for end repair, A-base addition, adaptor ligation, and library enrichment PCR were purchased from KAPA Biosciences in 96-reaction kits. In addition, during the post-enrichment SPRI cleanup, elution volume was reduced to 20 \( \mu \)L to maximize library concentration, and a vortexing step was added to maximize the amount of template eluted.

For Agilent capture, in-solution hybridization selection was performed as described by Fisher et al.,\textsuperscript{21} with the following exception: prior to hybridization, two normalized libraries were pooled together, yielding the same total volume and concentration specified in the publication. Following post-capture enrichment, libraries were quantified using quantitative PCR (kit purchased from KAPA Biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent’s Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized...
to 2 nM and pooled by equal volume using the Hamilton Starlet. Pools were then denatured using 0.1 N NaOH. Finally, denatured samples were diluted into strip tubes using the Hamilton Starlet.

For ICE capture, in-solution hybridization and capture were performed using the relevant components of Illumina's Rapid Capture Exome Kit and following the manufacturer's suggested protocol, with the following exceptions: first, all libraries within a library construction plate were pooled prior to hybridization. Second, the Midi plate from Illumina's Rapid Capture Exome Kit was replaced with a skirted PCR plate to facilitate automation. All hybridization and capture steps were automated on the Agilent Bravo liquid handling system. After post-capture enrichment, library pools were quantified using qPCR (automated assay on the Agilent Bravo), using a kit purchased from KAPA Biosystems with probes specific to the ends of the adapters. Based on qPCR quantification, libraries were normalized to 2 nM, then denatured using 0.1 N NaOH on the Hamilton Starlet. After denaturation, libraries were diluted to 20 pM using hybridization buffer purchased from Illumina.

Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using HiSeq v3 cluster chemistry and HiSeq 2000 or 2500 flowcells. Flowcells were sequenced on HiSeq 2000 or 2500 using v3 Sequencing-by-Synthesis chemistry, then analyzed using RTA v.1.12.4.2 or later. Each pool of whole-exome libraries was run on paired 76 bp runs, with an 8 base index sequencing read was performed to read molecular indices.

**Variant Calling and Annotation**

We performed joint variant calling for single nucleotide variants and indels across all samples in this cohort and ~6,500 control samples from the Exome Sequencing Project using GATK v3.4. Specifically, we used the HaplotypeCaller pipeline according to GATK best practices. Variant quality score recalibration (VQSR) was performed, and in the majority of analyses only “PASS” variants were investigated. The resultant variant call file (VCF) was annotated with Variant Effect Predictor v9.1,22 Loftee, dpNSFP-2.9.3,23 and MPC,24 A combination of GATK,25 bcftools, and Gemini26 was used to identify rare and predicted damaging variants. Specifically, variants with an allele count (AC) of ≤3 in gnomAD (a population cohort of 123,136 exomes) were considered rare, and variants annotated as loss of function (LoF: splice acceptor or donor variants, stop gained, stop lost, start lost, and frameshifts) or missense by VEP were considered potentially damaging. Other rare variants in previously described DBA-associated genes with other annotations or no annotation were investigated on a case by case basis. When family members had also undergone WES, variants were required to fit Mendelian inheritance (e.g., dominant for RP genes, hemizygous for GATA1 and TSR2). In each family, all rare and predicted damaging de novo or recessive mutations were also considered. In all cases, pathogenic variants reported by Clinvar as well as rare variants in genes known to cause other disorders of red cell production or bone marrow failure were also considered.27 All putative causal variants were manually inspected in IGV.28 Cohort quality control including the ancestry analysis, cryptic relatedness, and sex checks was performed using peddy.29 Specifically, PCA was performed on 1000 Genomes project samples for the overlap of variants measured in the DBA cohort with ~25,000 variants from samples in the 1000 Genomes project. DBA cohort samples were then projected onto these PCs, and ancestry in the DBA cohort was

<table>
<thead>
<tr>
<th>DBA Case Subjects</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>472</td>
<td>–</td>
</tr>
<tr>
<td>Families</td>
<td>63</td>
<td>13.3%</td>
</tr>
<tr>
<td>Unrelated</td>
<td>425</td>
<td>90.0%</td>
</tr>
<tr>
<td>WES (+ verification)</td>
<td>445</td>
<td>94.3%</td>
</tr>
<tr>
<td>Sanger only</td>
<td>27</td>
<td>5.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age at Sample Collection</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2 years</td>
<td>138</td>
<td>32.1%</td>
</tr>
<tr>
<td>2–10 years</td>
<td>143</td>
<td>33.3%</td>
</tr>
<tr>
<td>10–18 years</td>
<td>57</td>
<td>13.3%</td>
</tr>
<tr>
<td>18+ years</td>
<td>92</td>
<td>21.4%</td>
</tr>
<tr>
<td>Unknown</td>
<td>42</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sex</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>249</td>
<td>53.4%</td>
</tr>
<tr>
<td>Female</td>
<td>217</td>
<td>46.6%</td>
</tr>
<tr>
<td>Unknown</td>
<td>6</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical</td>
<td>269</td>
<td>85.9%</td>
</tr>
<tr>
<td>Mild</td>
<td>15</td>
<td>4.8%</td>
</tr>
<tr>
<td>Atypical</td>
<td>29</td>
<td>9.3%</td>
</tr>
<tr>
<td>Unknown</td>
<td>159</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Congenital Malformations</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>159</td>
<td>55.6%</td>
</tr>
<tr>
<td>One or more</td>
<td>127</td>
<td>44.4%</td>
</tr>
<tr>
<td>Head or face</td>
<td>53</td>
<td>18.5%</td>
</tr>
<tr>
<td>Limbs</td>
<td>44</td>
<td>15.4%</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>19</td>
<td>6.6%</td>
</tr>
<tr>
<td>Heart</td>
<td>41</td>
<td>14.3%</td>
</tr>
<tr>
<td>Short stature</td>
<td>25</td>
<td>8.7%</td>
</tr>
<tr>
<td>Unknown</td>
<td>186</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elevated eADA</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>79.4%</td>
</tr>
<tr>
<td>No</td>
<td>13</td>
<td>20.6%</td>
</tr>
<tr>
<td>Unknown</td>
<td>409</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steroid dependent</td>
<td>79</td>
<td>26.6%</td>
</tr>
<tr>
<td>No steroid trail yet</td>
<td>49</td>
<td>16.5%</td>
</tr>
<tr>
<td>Transfusion dependent</td>
<td>112</td>
<td>37.7%</td>
</tr>
<tr>
<td>Remission</td>
<td>33</td>
<td>11.1%</td>
</tr>
<tr>
<td>BMFT</td>
<td>12</td>
<td>4.0%</td>
</tr>
<tr>
<td>No treatment</td>
<td>12</td>
<td>4.0%</td>
</tr>
<tr>
<td>Unknown</td>
<td>175</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 1. Cohort Characteristics

<table>
<thead>
<tr>
<th>DBA Case Subjects</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>472</td>
<td>–</td>
</tr>
<tr>
<td>Families</td>
<td>63</td>
<td>13.3%</td>
</tr>
<tr>
<td>Unrelated</td>
<td>425</td>
<td>90.0%</td>
</tr>
<tr>
<td>WES (+ verification)</td>
<td>445</td>
<td>94.3%</td>
</tr>
<tr>
<td>Sanger only</td>
<td>27</td>
<td>5.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age at Sample Collection</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2 years</td>
<td>138</td>
<td>32.1%</td>
</tr>
<tr>
<td>2–10 years</td>
<td>143</td>
<td>33.3%</td>
</tr>
<tr>
<td>10–18 years</td>
<td>57</td>
<td>13.3%</td>
</tr>
<tr>
<td>18+ years</td>
<td>92</td>
<td>21.4%</td>
</tr>
<tr>
<td>Unknown</td>
<td>42</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sex</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>249</td>
<td>53.4%</td>
</tr>
<tr>
<td>Female</td>
<td>217</td>
<td>46.6%</td>
</tr>
<tr>
<td>Unknown</td>
<td>6</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical</td>
<td>269</td>
<td>85.9%</td>
</tr>
<tr>
<td>Mild</td>
<td>15</td>
<td>4.8%</td>
</tr>
<tr>
<td>Atypical</td>
<td>29</td>
<td>9.3%</td>
</tr>
<tr>
<td>Unknown</td>
<td>159</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Congenital Malformations</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>159</td>
<td>55.6%</td>
</tr>
<tr>
<td>One or more</td>
<td>127</td>
<td>44.4%</td>
</tr>
<tr>
<td>Head or face</td>
<td>53</td>
<td>18.5%</td>
</tr>
<tr>
<td>Limbs</td>
<td>44</td>
<td>15.4%</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>19</td>
<td>6.6%</td>
</tr>
<tr>
<td>Heart</td>
<td>41</td>
<td>14.3%</td>
</tr>
<tr>
<td>Short stature</td>
<td>25</td>
<td>8.7%</td>
</tr>
<tr>
<td>Unknown</td>
<td>186</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elevated eADA</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>79.4%</td>
</tr>
<tr>
<td>No</td>
<td>13</td>
<td>20.6%</td>
</tr>
<tr>
<td>Unknown</td>
<td>409</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment</th>
<th>no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steroid dependent</td>
<td>79</td>
<td>26.6%</td>
</tr>
<tr>
<td>No steroid trail yet</td>
<td>49</td>
<td>16.5%</td>
</tr>
<tr>
<td>Transfusion dependent</td>
<td>112</td>
<td>37.7%</td>
</tr>
<tr>
<td>Remission</td>
<td>33</td>
<td>11.1%</td>
</tr>
<tr>
<td>BMFT</td>
<td>12</td>
<td>4.0%</td>
</tr>
<tr>
<td>No treatment</td>
<td>12</td>
<td>4.0%</td>
</tr>
<tr>
<td>Unknown</td>
<td>175</td>
<td>–</td>
</tr>
</tbody>
</table>
predicted from the PC coordinates using a support vector machine trained on known ancestry labels from 1000 Genomes samples. Relatedness parameters were calculated (coefficient of relatedness, ibs0, ibs1, ibs2) using these variants and were compared to known relationships from the cohort pedigrees; cases that did not agree were manually validated and corrected. In all cases, sex checks (presence of heterozygous variants on the X chromosome) performed by peddy aligned with available cohort information.

**Targeted Sanger Sequencing**

The Primer3 program was used to design primers to amplify a fragment of 200–300 bp targeting a specific region of either exon or intron of the gene of interest. Polymerase chain reaction was performed using Dream Taq Polymerase (Life Science Technology, Cat# EP0701) and 30 μg of genomic DNA in a 15 μL reaction. The reaction was performed with an initial denaturation of 5 min at 94°C followed by 29 cycles of second denaturation at 94°C for 45 s, annealing at 57°C for 45 s, and extension at 72°C for 45 s. The final extension was performed at 72°C for 10 min. The PCR product was treated with the reagent ExoSAP-IT (USB) and submitted for Sanger sequencing to the Boston Children’s Hospital Molecular Genetics Core Facility. The resulted sequences were analyzed using Sequencher 4.8 software (Gene Codes) and compared with normal gene sequence provided through the UCSC Genome Browser.

**RNA-Seq and Splicing Analysis**

RNA was isolated using RNeasy kits (QIAGEN) according to the manufacturer’s instructions. 1–20 ng of RNA were forwarded to a modified Smart-seq2 protocol and after reverse transcription, 8–9 cycles of PCR were used to amplify transcriptome libraries. The Primer3 program was used to design primers to amplify a fragment of 200–300 bp targeting a specific region of either exon or intron of the gene of interest. Polymerase chain reaction was performed using Dream Taq Polymerase (Life Science Technology, Cat# EP0701) and 30 μg of genomic DNA in a 15 μL reaction. The reaction was performed with an initial denaturation of 5 min at 94°C followed by 29 cycles of second denaturation at 94°C for 45 s, annealing at 57°C for 45 s, and extension at 72°C for 45 s. The final extension was performed at 72°C for 10 min. The PCR product was treated with the reagent ExoSAP-IT (USB) and submitted for Sanger sequencing to the Boston Children’s Hospital Molecular Genetics Core Facility. The resulted sequences were analyzed using Sequencher 4.8 software (Gene Codes) and compared with normal gene sequence provided through the UCSC Genome Browser.

**Lymphoblastoid Cell Lines**

To generate lymphoblastoid cell lines from peripheral blood, Histopaque solution was used to isolate the buffy coat containing mononuclear cells. Mononuclear cells were transferred into a new tube and washed twice with PBS. Cells were resuspended into 2 mL complete RPMI 1640 containing 15% fetal bovine serum and 5% penicillin/streptomycin and glutamine. 2 mL of Epstein-Barr virus (EBV) solution was added, and cells were incubated at 37°C and 5% CO2 overnight. After adding 5 mL complete RPMI, cells were allowed to grow to confluence and maintained using the regular cell culture procedure. Epstein-Barr virus (EBV) was generated by growing B95-8 cells in RPMI complete until they were at a high cell concentration (1–2×10^9) for 12 to 14 days. Cells were centrifuged at 1,300 RPM for 10 min at 20°C. The supernatant (containing EBV virus) was passed through a PEB filter twice, aliquoted in 2 mL cryogenic vials, and stored at −80°C. This procedure was performed in accordance with the Boston Children’s Hospital’s Biosafety protocol.

**Copy Number Variant Identification and Validation**

Copy number variant (CNV) analysis was performed for the entire cohort using XHMM separately for ICE and Agilent exomes, as previously described. Specifically, XHMM takes as input a sample by exon read coverage matrix, performs principal component (PC) analysis, re-projects the matrix after removing PCs that explain a large proportion of the variance, normalizes the matrix (z-score), and then uses a hidden Markov model (HMM) to estimate copy number state. For known RP genes, candidate deletions were nominated either by (1) XHMM deletion calls or (2) manual investigation of outliers in the z-score distribution for each exon. When WES was performed in other family members, the inheritance of putative CNVs was also determined. Putative CNVs were validated using ddPCR. Specifically, primers and probes were designed to amplify exons with putative deletions. 50 ng of DNA per sample (at least one test and one control per reaction) were digested with a restriction enzyme, either Hind or HaeIII, and master mixes containing FAM targeted assays and control HEX RPP30 assays. Subsequently, plates were foil sealed, vortexed, and placed in an autodroplet generator (BioRad). Once the droplets were generated, plates were placed in thermal cycler C1000 Touch (BioRad) for DNA amplification. PCR was performed with an initial denaturing step at 95°C for 10 min, followed by 40 cycles of denaturing at 95°C for 30 s and annealing at 60°C for 1 min. Subsequently, enzyme deactivation was achieved by heating to 98°C for 10 min. Each PCR run included no-template controls and normal controls. The results of ddPCR were generated using QX200 Droplet Reader (BioRad) and analyzed using QuantaSoft Analysis Pro (BioRad).

**Segmental Duplication Analysis**

To investigate the copy number distribution of RPS17 in the human population, we used Genome STRIP to determine the copy number of this gene using whole-genome sequence data from the 1000 Genomes Project in 2,535 individuals of diverse ancestry. We first measured the copy number of the segmental duplication containing RPS17, specifying the coordinates of both copies of the segmental duplication (hg19 coordinates chr15:82629052–82829645 and chr15:83005382–83213987) to estimate the total copy number (which would be 4 for individuals homozygous for the hg19 reference haplotype that contains two copies of this segment). We further measured just a 5 kb segment directly at RPS17 (hg19 coordinates chr15:83205001–83210000 and chr15:82820658–82825658) to determine whether the gene itself was present in two diploid copies in individuals from the 1000 Genomes cohort. We also performed the same measurement
Penetration Analysis
Penetration analysis was performed as previously described in Minikel et al.41 with a few key modifications. Specifically, we use Bayes’ rule to obtain \( P(D|G) = \frac{P(D) \times P(G|D)}{P(G)} \), where \( P(D|G) \) is the penetrance for a specific genotype \( G \), \( P(D) \) is the lifetime risk of the disease \( D \) in a general population, \( P(G) \) is the proportion of individuals with \( DBA \) who have the specific genotype \( G \), and \( P(G|D) \) is the proportion of individuals in the general population who have the specific genotype. We can calculate \( P(D) \) as average lifetime \( \times \) DBA incidence = 80 years \( \times \) 7/1,000,000. We obtain an estimate for \( P(G|D) \) as \( \# \) individuals with \( \geq 1 \) allele in \( DBA \) cohort. Similarly, we can obtain an estimate for \( P(G) \) as \( \# \) individuals with \( \geq 1 \) allele in \( gnomAD \) + 1, where we add 1 (estimated integer of DBA-affected case subjects in a population of size 121,136) to the proportion of individuals with the specific genotype in \( gnomAD \), since \( gnomAD \) is not perfectly representative of the general population and most or all potential \( DBA \)-affected case subjects are likely to have been removed. This allows us to plug in to calculate a point estimate for \( P(D|G) \) as \( \min(1, \frac{P(D) \times P(G|D)}{P(G)}) \). Similarly, we can quantify the spread in this estimate using 95% Wilson confidence intervals of a binomial distribution (also known as score intervals). We note that by adding 1 to the denominator this could potentially result in a slightly lower and more conservative estimate of penetrance. Since the majority of variants identified were singletons and we are primarily interested in inference at the gene and variant type level, we collapsed variants by predicted effects (LoF, missense) and gene in order to obtain more robust estimates. A max total allele count \( (AC) \) of 12 across the combined set of \( DBA \) and \( gnomAD \) exomes was used as a filter, since a few variants reached higher prevalence in \( DBA \). The penetrance of the mutation identified as polymorphic from the \( DBA \) genes database (\( RPL5; \) c.418G>A) was estimated using the same formula, and \( P(G|D) \) was estimated using the \( DBA \) cohort in this study (one individual was observed to have the A allele).

Structural Analysis
The cryo-EM structure of the human 80S ribosome and P-stalk proteins from the cryo-EM structure of the yeast 80S ribosome were used to create a hybrid 80S structural model shown in Figure 1E. Structural superposition, analyses, and figures were rendered using PyMOL.44

Gene Burden Analyses
Gene-based burden testing was performed using TRAPD which employs a one-sided Fisher’s exact test of the 2x2 table of genotype counts (genotype present or genotype absent) between case subjects (407 unrelated case subjects from the \( DBA \) cohort) and control subjects (\( gnomAD \)). \( gnomAD \) is an aggregation database of exome sequencing from 123,136 individuals who are not known to have a severe Mendelian condition. Counts under the dominant model were generated for \( DBA \) by counting the number of individuals who carry at least one qualifying variant in each gene and for \( gnomAD \) by summing the allele counts for qualifying variants in each gene. For the recessive model, counts
in DBA were generated by counting the number of individuals who carry two or more qualifying variants in a gene or who are homozygous for a qualifying variant. For the recessive model in gnomAD, the number of individuals who carry a homozygous variant was added to a predicted number of compound heterozygous variant carriers. The predicted number of compound heterozygous variant carriers was calculated by squaring the total heterozygous variant carrier rate in each gene and multiplying by the total sample size. p values < 2.5 × 10⁻⁶ were considered significant (of 0.05 corrected for testing ≈ 20,000 genes). Predicted damaging missense mutations were identified using PolyPhen-2. Several steps were taken to match variant call set quality since variants were not jointly called for the case and control subjects. First, read depth was computed in each cohort separately and only sites where the read depth was >10 in each cohort were removed. Second, sites present in low complexity regions were removed. Third, rare synonymous variant burden testing was performed for different variant quality score recalibration (VQSR) combinations until the −log₁₀ p values from the Fisher’s exact test followed the expected distribution. Specifically, inclusion of the top 85% of VQSR variants from the DBA cohort and the top 95% of VQSR variants from the gnomAD cohort resulted in the best fit (erring slightly to be more conservative than not). For gene set enrichment tests, autosomal-dominant Mendelian control genes were taken as the union of two previously reported studies.

Statistical Analyses
In order to test for differences in outcome (e.g., congenital abnormalities, treatment outcomes), a Fisher’s exact test was performed on the genotype by phenotype count matrix and p values were calculated from 100,000 Monte Carlo simulations. The type of mutation (e.g., LoF or missense) was not separately investigated, since this is confounded by the exact gene implicated, although gene-phenotype associations were secondarily validated after removing missense mutations. For outcomes of interest, 95% binomial confidence intervals are reported in addition to the point estimates. Precision recall curves for classification of RPS19 missense mutations as belonging to the DBA cohort or to the gnomAD cohort based upon missense pathogenicity predictive methods were calculated using the R package pROC. Other enrichment tests (e.g., splicing position) were calculated using Fisher’s exact tests on 2×2 tables of variant counts. Power analysis for burden testing was performed using the power.fisher.test function in the R statmod package with at least 10,000 simulations.

Results
Overall Yield and Spectrum of Rare Variants in the DBA Cohort
We assembled a cohort of 472 individuals, 47 of whom were related to at least one other individual, with a likely diagnosis of DBA of predominantly European descent (76%) from DBA registries and clinicians over the course of 20 years and performed WES on 94% (Figure 1A). In an attempt to comprehensively identify or verify causal mutations, we investigated rare LoF and missense mutations in genes known to harbor pathogenic DBA mutations (DBA-associated genes), called CNVs using WES coverage estimates, obtained and analyzed RNA-seq data from patient samples to determine the pathogenicity of cryptic splice mutations, and performed gene burden analyses to nominate new genes. Combining all approaches (Material and Methods), we identified putative causal mutations for the observed anemia in 78% (367/472) of case subjects (Figure 1B, Tables S2 and S3). The majority of these mutations were in one of the 19 previously known genes known to harbor pathogenic DBA mutations (330/472, 70%) and were primarily rare (gnomAD AC ≤ 3) loss-of-function (LoF) or missense alleles identified from WES. Twenty-seven case subjects did not undergo WES due to limitations in available material but had known rare LoF or missense alleles identified by Sanger sequencing. Most putative causal mutations were typical LoF alleles or disrupted canonical mRNA splice sites (Figure 1C). In agreement with previous reports, RPS19, RPL5, RPS26, and RPL11 were the most frequently mutated RP genes (Figures 1B and 1E). The majority of mutations were unique, with 80% of mutations observed in not more than one unrelated case (Figure 1D). Eleven (2%) case subjects had two distinct rare and putatively damaging RP gene mutations. Sanger sequencing validated 100% of putative causal mutations identified. However, a small but considerable number of DBA gene mutations (7/472, 1.4%) were identified from targeted Sanger sequencing of the genes known to harbor pathogenic DBA mutations but were not found in the initial variant calls from WES (Table S4). While a few of these mutations were in genes duplicated in the hg19 genome build (RPS17) or in regions with low coverage (start site of RPS24), the majority were long and/or complex indels. Thus, although WES is highly accurate, specific classes of clinically relevant LoF mutations, such as medium-sized indels, can be missed, and our results suggest a benefit to performing follow-up targeted Sanger or long-read sequencing when WES does not return a high-confidence causal mutation. In total, among the 335 individuals without an RP gene CNV, 71 individuals (21%) harbored 1 of 56 novel variants that met our criteria for being a putative causal mutation, including 40 LoF, 6 canonical splice site, 2 extended splice region, 7 missense, and 1 inframe insertion.

Extended and Cryptic Splice Site Mutations in Genes Known to Harbor Pathogenic DBA Mutations
Mutations that alter splicing but that lie outside canonical splice donor or acceptor sites, including deep intronic variants, have recently been shown to account for a substantial fraction of Mendelian disease cases with previously unknown pathogenic variants. Popular annotation tools, such as Variant Effect Predictor (VEP) and SnpEff, define mutations that disrupt only the first two (GT/U) or last two (AG) intronic bases as canonical splice site mutations. Although WES can detect mutations only in sequences captured by exome baits and thus misses the majority of intronic bases, it can detect mutations in proximal splice sites. We observe not only an enrichment for canonical splice region mutations but also a substantial
increase in “extended” splice region mutations in known DBA-associated genes in our cohort compared to 123,136 population control subjects from gnomAD (Figure S1). While these mutations predominately affect the third base of the extended consensus splice acceptor or donor site (proband 1; Figure 2C), we identified a small number of rare mutations further from the exon-intron junction that are not typically considered. For example, we identified a mutation eight bases upstream of RPS26 that are not typically considered. For example, we identified a mutation eight bases upstream of RPS26 that was absent from gnomAD (proband 3, Figure 2C). This mutation is predicted to create a novel consensus acceptor site (TAT> TAG) that would likely result in a frameshift due to the inclusion of seven additional nucleotides into the RPS26 transcript.

Since RP genes are ubiquitously expressed, we reasoned that performing RNA sequencing (RNA-seq) in cell lines derived from affected individuals would help us to determine whether these extended splice region mutations were in fact splice disrupting. Therefore, for five healthy control subjects and eight case subjects with extended splice regions mutations, we created lymphoblastoid cell lines (LCLs) and performed RNA-seq. For six of the eight case subjects, we observed aberrant splicing of the RP genes and/or decreased mRNA expression (Figures 2A–2C and S2). In several cases, a mutation at the third base of a splice donor or acceptor site resulted in exon skipping (probands 1 and 5, Figures 2A–2C and S2A). Interestingly, for two unique RPS26 mutations that were each eight bases upstream of a different coding exon and created potential splice acceptor sites, we observed novel exon extensions (probands 2 and 3, Figures 2B and 2C). In the case mentioned above, a novel acceptor site was created and used, resulting in a frameshift (proband 3, Figure 2C). In another case, the presumed mutant acceptor site was not faithfully used, again resulting in the introduction of a frameshift to the transcript (proband 2, Figure 2C). The acceptor mutation in one individual was so severe that only limited splicing seemed to occur on the mutated transcript, and a substantial proportion of polyadenylated transcripts appeared to have intron retention (proband 7, Figures S2C and S2D). Together, these results suggest that a proportion of case subjects lacking a typical RP gene mutation may instead harbor cryptic splicing mutations or mutations with post-transcriptional effects in one of the 19 currently known DBA-associated genes. As the size of population-based WGS databases grows, identifying such cryptic mutations should become increasingly feasible with WGS. Furthermore, given the ubiquitous expression of RP genes, RNA-seq of affected individual-derived LCLs or fibroblasts could prove to be a relatively straightforward and fruitful strategy for identifying the functional impact on splicing or transcript expression from causal mutations missed by WES.

Identification of a Null Mutation in the 3’ UTR of RPS26

We next extended our analysis to rare mutations that did not appear to be canonical LoF, missense, or splice region mutations in known RP genes. Interestingly, we identified a mutation in the 3’ UTR of RPS26 that was also absent from gnomAD. This mutation was predicted to completely disrupt the polyadenylation signal (PAS) by changing the consensus motif AA(T/U)AAA to AAGAAA (proband 4, Figure 2C). To test whether this was in fact the case, we created a LCL from this individual and performed RNA-seq. We found that transcription continued approximately 700 bases past the typical polyadenylation site, drastically increasing the size of the 3’ UTR (Figure 2B). Furthermore, mRNA levels of RPS26 were significantly reduced, although global mRNA profiles were largely...
similar (Figure S2E). Although not tested here, it is likely that RPS26 transcripts with the long mutant 3' UTRs are less stable and targeted by miRNAs or RNA-binding proteins, resulting in reduced mRNA levels.

**Copy Number Variants in Genes Known to Harbor Pathogenic DBA Mutations**

In addition to missense or LoF mutations, smaller studies have estimated that 15%–20% of DBA-affected case subjects are due to a partial or full deletion of one copy of an RP gene.54,55 Although an imperfect approach, copy number variants (CNVs) can be identified as differences in coverage across regions ascertained by WES.36,37 Thus, we performed WES-based CNV calling (Material and Methods) and identified 79 putative deletions in known DBA-associated genes plus other RP genes (Table S5). To verify a subset of these deletions, we used digital droplet PCR (ddPCR) to test 13 of the most commonly deleted exons, representing 7 genes and 44 case subjects. Reflective of the fact that these putative deletions were carefully preselected as high confidence CNVs, 29 (66%) could be verified by ddPCR, 2 CNVs were identified by array CGH, and 2 very high confidence RPS17 microdeletions were considered diagnostic although they were not validated due to limited DNA samples (Figure 3 and Table S6). RPS17 was the most frequently deleted gene (11 case subjects, 2 verified as true de novo), predominately occurring as part of the well-described 15q25.2 microdeletion56 (Figure 3D). Notably, in hg19, RPS17 is annotated as a duplicated gene (RPS17 and RPS17L) and previous studies have considered whether a loss of one copy (out of four) was sufficient to result in DBA.55 Here, using empirical estimates of copy number from WGS for this region (Figure S3), we determined that a duplication event is not supported and that there is in fact only a single copy of RPS17 (indeed, this appears to have been resolved in hg38). Given our overall success here and in other studies of rare blood diseases,37 we recommend that WES-based CNV calling should become a standard part of clinical WES analysis.

**Prevalence and Penetrance of Mutations in Genes Known to Harbor Pathogenic DBA Mutations**

Although LoF and missense mutations occur far less frequently than expected in the majority of genes known to harbor pathogenic DBA mutations (Table S7),14 the exact prevalence and penetrance of different allele frequency (AF) classes of DBA gene mutations have not been systematically investigated. We first investigated whether the class of more common but still rare (0.005% to 1%) missense mutations in DBA-associated genes was enriched in case subjects compared to gnomAD, but observed a non-significant odds ratio of $\approx 1$ (Figure S4A). Since analyses of higher frequency variants may be confounded by unaccounted population structure, we used a set of unrelated dominant Mendelian genes as a control. We observed a larger enrichment for control genes than for DBA-associated genes, which suggests that, if anything, our results were biased against the null of no association between these variants and DBA (Figure S4A; Material and Methods).

There have been several reports of incomplete penetrance or variable expressivity for rare RP gene mutations.57–59 Therefore, we set out to investigate the relative penetrance of rare DBA-associated gene mutations.11 Since no mutation was present in the DBA cohort at an allele count (AC) higher than 8 (14 including related individuals), we grouped mutations by both gene and type (LoF or missense). For the three most frequently mutated DBA-associated genes (RPS19, RPL5, and RPS26), we found that LoF mutations demonstrate nearly complete
penetrance (Figure 4A). When we grouped LoF mutations in the less commonly mutated DBA-associated genes, we similarly found that these mutations were also highly penetrant, although the point estimate was lower than for the more common DBA-associated genes (Figure 4A).

This potentially suggests some variable expressivity of known DBA gene mutations, but this observation is confounded by the fact that not all predicted LoF mutations will cause a true loss of protein production. Since the majority of missense mutations were in RPS19 (42/73, 58%), we investigated the penetrance of this class and found that rare missense mutations, in aggregate, were far less penetrant (6%) than rare LoF mutations (Figure 4A). To determine whether the missense mutations in the cohort were predicted to be more damaging than similarly rare missense mutations in gnomAD, we annotated all variants using Envision, which, unlike most predictors, is not trained on gnomAD or databases of Mendelian mutations. Using Envision scores, we observed that damaging RPS19 missense mutations were predicted to have higher penetrance (22%, Figure 4A) and that the majority of RPS19 missense mutations in our DBA cohort were more damaging than those in gnomAD (Figure 4B). However, we caution against over-interpretation, as not even predictive algorithms that were in part trained on RP gene mutations and/or gnomAD could perfectly separate DBA missense mutations from gnomAD (auPRC range of 0.53 to 0.69).

We next investigated the specific impacts of RPS19 missense mutations, using both our cohort and a curated database of pathogenic DBA mutations (DBAgenes). We observed three distinct types of mutations in our cohort. First, 48% of mutations changed a hydrophobic amino acid (aa) to a non-hydrophobic amino acid. Second, = 18% of mutations changed a non-special aa to a special aa, such as proline. Third, 13% of mutations changed the smallest aa, glycine, to a much larger aa. These three types accounted for 78% of all DBA mutations but for only 25% of gnomAD mutations (p = 0.006).

Structurally, we observed that 85% of RPS19 missense mutation-carrying individuals in our cohort contained a mutation within exons encoding the four core α helices (1, 2, 4, and 5; Figure 4C). Given that the locations of these mutations along the mRNA transcript were not fully independent from those variants observed in gnomAD, we investigated whether DBA mutations were more likely to affect specific structural elements of the RPS19 protein in the context of the fully assembled human ribosome (Figure 4D). A high-resolution ribosome structure shows that four major α helices form a hydrophobic core to stabilize RPS19 (Figure 4E). Consistent with a previous report and our observation that DBA mutations often disrupt hydrophobic amino acids, we determined that ≈ 50% of mutations would destabilize this hydrophobic core (Table S8). Furthermore, RPS19 stabilizes two long hairpin ribosomal RNA (rRNA) regions at the head of the small subunit (Figure 4D). These interactions would be disrupted by ≈ 43%, suggesting that the second largest class of missense mutations in RPS19 would disrupt interactions between RPS19 and rRNA, consistent with a previous hypothesis.

Finally, motivated by a previous report that re-assessed the penetrance and pathogenicity of variants associated
with Mendelian disease in public databases,41 we investigated the frequency of reported variants from the DBA-genes database63 in the gnomAD population control. Importantly, 202/203 of the reported mutations in this database had 1 or fewer allele counts in gnomAD, consistent with our study as well as the low incidence and phenotypic severity of DBA. However, one missense variant (RPL5: c.418G>A, chr1:93301840:G>A) was relatively more common and was observed in 27 individuals in gnomAD, indicating that this variant either was not pathogenic or has low penetrance (point estimate (2.5%–97.5% CI); 0.0%, 0.0%–4.7%). Overall, our results indicate that rare LoF mutations in RP genes almost always result in DBA, whereas missense or more common mutations require increased scrutiny. Therefore, it is important for clinicians or families. In agreement with previous studies on smaller cohorts (including a subset of this cohort),44,46,60–62 we observed significant differences in the presence of congenital malformations among individuals with mutations in different RP genes (Figure 5A). In fact, the majority of individuals with RPL5 (point estimate (2.5%–97.5% CI); 83%, 67%–93%) or RPL11 (73%, 50%–88%) mutations had one or more congenital malformations, in contrast to individuals with RPS19 mutations where only 34% (24%–47%) had any congenital malformation. In addition to observing significant associations between RP gene and congenital malformations affecting the head and face, limbs, stature, and genitourinary system, we also observed a significant association with the presence of congenital heart disease, which has previously been underappreciated65,67 (Figures 5B and S5). Leveraging the size of our cohort to make robust estimates, we conclude that 22% (12%–35%) and 13% (4%–31%) of individuals with RPL5 and RPL11 mutations, respectively, present with cardiac abnormalities, which is in stark contrast to the 4% (1%–9%) and 7% (2%–20%) of individuals with RPS19 and RPS26 mutations.

We next investigated whether there were differences in treatment requirements between RP genes for the primary condition of anemia, and we observed a significant association (Figure 5C). However, this did not appear to be due to differences in transfusion or corticosteroid treatment dependence, which account for approximately 65% of individuals (Table 1). Instead, we observed a difference between RP genes only for the proportion of individuals that were in remission (p = 0.001) (Figure S5E). This appeared to be driven by the observation that 36% (14%–64%) and 29% (12%–52%) of individuals with RPS24 and RPL11 mutations were in remission and currently required no treatment, whereas only 8% (4%–17%) and 5% (1%–20%) of RPS19 and RPL5 individuals were in remission. After removing these individuals, the original association between RP gene and treatment requirement was no longer significant (p = 0.14), indicating that the major difference in treatment requirement between individuals with disparately mutated RP genes is the likelihood of remission.

Finally, we investigated whether there were differences in erythrocyte adenosine deaminase (eADA) levels, since elevated eADA is a useful diagnostic biomarker in DBA.68,69 eADA measurement information was available for only 63 individuals and 79% were observed to have an elevated eADA, consistent with recent studies.68,69 Although these studies reported little to no differences in eADA levels between RP genes, we observe a significant

---

**Figure 5. Phenotypic Associations**

Differences in (A) presence of ≥1 congenital abnormality, (B) heart malformations, (C) treatment requirements, and (D) eADA levels were observed between different RP genes. Differences remained significant after removing missense mutations. A Fisher’s exact test was used to test the hypothesis that there were differences in proportion of the outcome between RP genes. (D) The association between RP gene and treatment requirements appeared to primarily be due to differences in remission, and this association was no longer significant after removing individuals who went into remission.

---

**Phenotype-Genotype Associations in Known DBA-Associated Genes**

Although detailed phenotypic information was unavailable for a portion of the cohort (Table 1), we were nonetheless able to investigate phenotypic differences between individuals with disparate RP gene mutations. This information was primarily obtained by report from referring clinicians or families. In agreement with previous studies on smaller cohorts (including a subset of this cohort),65,66 we observed significant differences in the presence of congenital malformations among individuals with mutations in different RP genes (Figure 5A). In fact, the majority of individuals with RPL5 (point estimate (2.5%–97.5% CI); 83%, 67%–93%) or RPL11 (73%, 50%–88%) mutations had one or more congenital malformations, in contrast to individuals with RPS19 mutations where only 34% (24%–47%) had any congenital malformation. In addition to observing significant associations between RP gene and congenital malformations affecting the head and face, limbs, stature, and genitourinary system, we also observed a significant association with the presence of congenital heart disease, which has previously been underappreciated65,67 (Figures 5B and S5). Leveraging the size of our cohort to make robust estimates, we conclude that 22% (12%–35%) and 13% (4%–31%) of individuals with RPL5 and RPL11 mutations, respectively, present with cardiac abnormalities, which is in stark contrast to the 4% (1%–9%) and 7% (2%–20%) of individuals with RPS19 and RPS26 mutations.

We next investigated whether there were differences in treatment requirements between RP genes for the primary condition of anemia, and we observed a significant association (Figure 5C). However, this did not appear to be due to differences in transfusion or corticosteroid treatment dependence, which account for approximately 65% of individuals (Table 1). Instead, we observed a difference between RP genes only for the proportion of individuals that were in remission (p = 0.001) (Figure S5E). This appeared to be driven by the observation that 36% (14%–64%) and 29% (12%–52%) of individuals with RPS24 and RPL11 mutations were in remission and currently required no treatment, whereas only 8% (4%–17%) and 5% (1%–20%) of RPS19 and RPL5 individuals were in remission. After removing these individuals, the original association between RP gene and treatment requirement was no longer significant (p = 0.14), indicating that the major difference in treatment requirement between individuals with disparately mutated RP genes is the likelihood of remission.

Finally, we investigated whether there were differences in erythrocyte adenosine deaminase (eADA) levels, since elevated eADA is a useful diagnostic biomarker in DBA.68,69 eADA measurement information was available for only 63 individuals and 79% were observed to have an elevated eADA, consistent with recent studies.68,69 Although these studies reported little to no differences in eADA levels between RP genes, we observe a significant
association where RPS19 and RPS24 individuals appear less likely to have elevated eADA (Figure 5D). However, we caution that larger studies are required to determine whether this observation is robust. Overall, these findings highlight the differences in clinical features due to disparate RP gene mutations.

**Putative Pathogenic Mutations in Additional RP Genes**

We next investigated whether we could identify additional RP genes involved in DBA. To identify putative causal mutations, we similarly searched for rare (gnomAD AC ≤ 3) LoF and missense mutations in the remaining 60 RP genes without previously reported DBA mutations. A total of 9 mutations (7 unique) involving 7 previously unreported RP genes were identified (Table 2). These 7 RP genes are extremely intolerant to LoF mutation (minimum pLI = 0.73, maximum o/e = 0.12), similar to nearly all other previously reported DBA-associated genes (Table S7).

Two of the identified mutations were in splice regions and one altered a start codon. The other mutations were missense and were predicted by multiple algorithms to have damaging effects, whereas another mutation was in an RP gene encoded on the X chromosome in a male individual. Although we do not explicitly validate any of these putative pathogenic mutations here, it is likely with additional follow-up functional and genetic studies of DBA-affected individuals that additional evidence for a causal role of these genes may be established.

**Exome-wide Significant Genes in DBA**

Having extensively characterized mutations in the known DBA-associated genes and other RP genes, we sought to identify novel genes associated with the clinical features characteristic of DBA by performing gene burden tests between unrelated individuals in our cohort and gnomAD control subjects (a cohort presumably depleted of rare pediatric diseases). We first carefully adjusted the variant quality thresholds between the case and control subjects such that no genes were more enriched for rare (max DBA + gnomAD AC ≤ 3 or 6) synonymous mutations in the DBA cohort than expected (Figure 5A; Material and Methods). Restricting to rare LoF and damaging missense mutations with dominant inheritance, we identified RPS19 (30% prevalence), RPL5 (12%), RPS26 (9%), RPL11 (7%), and RPS10 (1%) as significantly associated with DBA at an exome-wide significant threshold (p = 0.05 / 20,000 = 2.5 × 10^{-6}) (Figures 6B and 6C). If we additionally included all missense mutations, we observed a sub-threshold association for RPL35A (2%; p = 0.00001) with DBA. Together, mutations in these 5 genes account for 59% of DBA-affected individuals in the cohort. However, we did not observe strong associations for the more prevalent genes RPS24 (3%) and RPS17 (3%), primarily because a large proportion of mutations in these genes were large deletions (we attempted to perform a CNV burden analysis but were unable to properly control inflation). Among all other non-RP genes, we observed SEH1L, HNRPCC, and ERCC1 as significantly associated with DBA at at least one test, but upon manual inspection the mutation calls were determined to be due to spurious alignments. Since we are both theoretically (Figure S5) and empirically (Figures 6B–6D) well powered to detect genes containing mutations of clear effects (such as LoF or damaging missense), we conclude that it is unlikely that dominant mutations in any single unknown gene that are detectable by WES are causal for more than 5% of case subjects of DBA with unknown genetic etiology.

Although we did not identify any new genes associated with DBA at the exome-wide significant level for dominant inheritance, we performed similar gene burden tests for recessive inheritance. For rare (max DBA + gnomAD AC ≤ 20) LoF and damaging missense mutations with recessive inheritance, we identified one exome-wide significant gene, which was CECRI (MIM: 607575)
cases that either phenocopied or caused DBA. We conservatively identified 30 (6%) rare and predicted damaging genotypes in known or suspected red cell/bone marrow failure disorder genes that were non-RP genes (Table 3). Although the majority of these were in CECE1 (9) or were mitochondrial deletions indicative of Pearson’s syndrome (7 [MIM: 557000]), as has been previously reported, 76 we identified several genes of interest that were mutated in a small number of case subjects. First, our cohort contained four individuals with GATA1 mutations, 11,16 two related individuals with a shared TSR2 mutation, 13,14 and one individual with a rare EPO missense mutation, 27 each of which we have previously reported. We have shown that TSR2, which is an RPS26 chaperone, has a critical role in ensuring adequate ribosome levels in hematopoietic progenitors and that altered GATA1 translation occurs due to RP haploinsufficiency, 14,16 suggesting that these mutations result in DBA through a common pathway. On the other hand, the missense mutation in TSR2, which there were other identifiable (Figure 6D). In total, we identified nine individuals with recessive or compound heterozygous missense or LoF mutations in CECE1, including two independent families in which recessive inheritance tracks with DBA status (Table 3). Although bi-allelic mutations in CECE1 (that result in deficiency of ADA2) were initially associated with vasculitis, several recent reports have identified similar mutations in less than a handful of individuals diagnosed with DBA or pure red cell aplasia. 70–73 Here, each of the individuals presented with severe normocytic or microcytic anemia and bone marrow erythroid hypoplasia in infancy without any additional physical abnormalities. When we performed pre-rRNA maturation assays on whole blood from two unrelated CECE1 individuals, we did not observe abnormal RNA maturation that is typical of RP gene DBA. 74 Preliminary evidence suggests that even though ADA2 encodes an adenosine deaminase, similar to ADA, these individuals were not observed to have elevated eADA unlike the majority (85%) of DBA-affected individuals. However, it is important to note that hematopoietic stem cell transplant appears curative in such individuals, 70,75 suggesting that this disorder may emerge due to a hematopoietic intrinsic defect, although not necessarily intrinsic to the erythroid compartment itself. Overall, our data suggest that individuals presenting with DBA should be screened for CECE1 mutations in addition to other known DBA-associated genes and that this is a condition that must be considered in any individual presenting with hypoplastic anemia.

Phenocopies, Misdiagnoses, and Non-RP Gene Mutations

Although CECE1 was the only non-RP gene that was associated with a diagnosis of DBA at exome-wide significance, we investigated the extent to which there were other identifiable
that lacked a typical DBA gene mutation (Table 3). First, we identified two unrelated individuals with recessive LoF mutations (one new, one known) in SLC25A38 (MIM: 610819), a known causal gene for congenital sideroblastic anemia (CSA [MIM: 205950]).77 Given the heterogeneous presentation of CSA, it is possible that the characteristic ring sideroblasts were initially limited or absent in samples obtained at diagnosis or that they were simply missed. In another individual, we identified compound heterozygous LoF mutations in PUS1 (MIM: 608109), which encodes for pseudouridine synthase 1. Only a handful of individuals with PUS1 mutations have been reported, but the majority presented with sideroblastic anemia, mitochondrial myopathy, and other dysmorphic features (MIM: 608109).78–80 In another individual, we observed a novel recessive LoF mutation in MYSM1 (MIM: 612176). Mutations in this gene have been previously described in only a handful of individuals,81,82 each presenting with transfusion-dependent refractory anemia in early childhood in addition to other cytopenias (MIM: 618116).83 Finally, we identified one individual with a recessive LoF mutation in NHEJ1 (MIM: 611290). Several individuals with similar mutations have been described as having immunodeficiency and dysmorphic faces, and in about half of them, anemia and thrombocytopenia (MIM: 611291).84 These findings suggest that DBA may be misdiagnosed in a small but important subset of individuals who in fact have one of a number of rare diseases in which hypoplastic anemia is a component of the phenotype.

Discussion

This work provides a systematic study on the approaches that can be used and the difficulties encountered when attempting to comprehensively define causal genetic lesions involved in a single Mendelian disorder. Even though we were conservative in assigning “causality,” we had a high genetic diagnosis rate of 78%, which is higher than other large reports on DBA cohorts and also higher than for most other Mendelian diseases. We achieved this high yet conservative diagnostic rate by leveraging large-scale population genetic databases (gnomAD) to remove “common” variants down to an allele frequency of 0.003%, by using multiple modern predictive algorithms when assigning pathogenicity, and by carefully investigating less-well-annotated mutations in or near known genes. To gain extra information from WES, we also used coverage information to nominate CNVs, 31 of which we orthogonally validated. Given the ubiquitous expression of RP gene mutations, we applied RNA-seq to LCLs derived from individuals with DBA and unambiguously validated six extended or cryptic splicing mutations and a 3' UTR mutation in RPS26. Finally, we did note a small, but significant, improvement in our genetic diagnostic rate by

<table>
<thead>
<tr>
<th>Gene</th>
<th>Inheritance</th>
<th># Affected</th>
<th>Mutation 1</th>
<th>Mutation 2</th>
<th>Type</th>
<th>gnomAD AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson (MT)</td>
<td>mitochondrial</td>
<td>7</td>
<td>mitochondrial deletion</td>
<td>–</td>
<td>deletion</td>
<td>—</td>
</tr>
<tr>
<td>CECR1</td>
<td>recessive</td>
<td>1</td>
<td>chr22:17662818:A&gt;T</td>
<td>chr22:17662818:A&gt;T</td>
<td>missense</td>
<td>0</td>
</tr>
<tr>
<td>CECR1</td>
<td>recessive</td>
<td>1</td>
<td>chr22:17662468:T&gt;A</td>
<td>chr22:17662468:T&gt;A</td>
<td>splice acceptor</td>
<td>0</td>
</tr>
<tr>
<td>CECR1</td>
<td>recessive</td>
<td>2</td>
<td>chr22:17669238:C&gt;T</td>
<td>chr22:17669238:C&gt;T</td>
<td>missense</td>
<td>6</td>
</tr>
<tr>
<td>CECR1</td>
<td>compound heterozygous</td>
<td>1</td>
<td>chr22:17690294:GC&gt;G</td>
<td>chr22:17690294:GC&gt;G</td>
<td>frameshift, missense</td>
<td>0, 2</td>
</tr>
<tr>
<td>CECR1</td>
<td>recessive</td>
<td>3</td>
<td>chr22:17662748:GTACGCT&gt;G</td>
<td>chr22:17662748:GTACGCT&gt;G</td>
<td>frameshift</td>
<td>1</td>
</tr>
<tr>
<td>CECR1</td>
<td>recessive</td>
<td>1</td>
<td>chr22:17690423:GC&gt;G</td>
<td>chr22:17690423:GC&gt;G</td>
<td>frameshift</td>
<td>0</td>
</tr>
<tr>
<td>GATA1</td>
<td>hemizygous</td>
<td>2</td>
<td>chrX:48649736:G&gt;C</td>
<td>–</td>
<td>missense</td>
<td>0</td>
</tr>
<tr>
<td>GATA1</td>
<td>hemizygous</td>
<td>1</td>
<td>chrX:48649735:AG&gt;A</td>
<td>–</td>
<td>frameshift</td>
<td>0</td>
</tr>
<tr>
<td>GATA1</td>
<td>hemizygous</td>
<td>1</td>
<td>chrX:48649518:T&gt;C</td>
<td>–</td>
<td>start lost</td>
<td>0</td>
</tr>
<tr>
<td>SLC25A38</td>
<td>recessive</td>
<td>1</td>
<td>chr3:39431108:G&gt;T</td>
<td>chr3:39431108:G&gt;T</td>
<td>splice donor</td>
<td>0</td>
</tr>
<tr>
<td>SLC25A38</td>
<td>recessive</td>
<td>1</td>
<td>chr3:39436065:A&gt;T</td>
<td>chr3:39436065:A&gt;T</td>
<td>stop gained</td>
<td>0</td>
</tr>
<tr>
<td>TSR2</td>
<td>hemizygous</td>
<td>2</td>
<td>chrX:54469851:A&gt;G</td>
<td>chrX:54469851:A&gt;G</td>
<td>missense</td>
<td>0</td>
</tr>
<tr>
<td>PUS1</td>
<td>compound heterozygous</td>
<td>1</td>
<td>chr12:132414269:T&gt;G</td>
<td>chr12:132414269:T&gt;G</td>
<td>stop gained, frameshift</td>
<td>0, 0</td>
</tr>
<tr>
<td>EPO</td>
<td>recessive</td>
<td>1</td>
<td>chr7:100320704:G&gt;A</td>
<td>chr7:100320704:G&gt;A</td>
<td>missense</td>
<td>1</td>
</tr>
<tr>
<td>NHEJ1</td>
<td>recessive</td>
<td>1</td>
<td>chr2:220011458:G&gt;A</td>
<td>chr2:220011458:G&gt;A</td>
<td>stop gained</td>
<td>2</td>
</tr>
<tr>
<td>MYSM1</td>
<td>recessive</td>
<td>1</td>
<td>chr1:59141211:G&gt;A</td>
<td>chr1:59141211:G&gt;A</td>
<td>stop gained</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviation: AC, allele count
Sanger sequencing of known genes when WES was inconclusive, as Sanger sequencing can identify medium length and complex indels, a class of variants that is currently inadequately detected by WES approaches.

Although the phenotypic expression of DBA is largely homogeneous, we observed that 6% of case subjects lacked typical mutations and instead harbored mutations that appeared to result in a phenocopy of DBA. Screening for causal mutations in DBA has typically been done using targeted Sanger sequencing of a handful of RP genes, but our work suggests that WES or WGS offers a substantial improvement. For example, we identified recessive CECR1 mutations in nine individuals in our cohort, highlighting the importance of screening for CECR1 mutations in individuals with a clinical DBA diagnosis. Although we were well powered to identify novel genes harboring rare LoF or damaging missense alleles, we did not identify any novel causal genes at an exome-wide significant level for DBA via gene burden testing. This suggests that larger sample sizes are needed to identify additional genes that harbor causal mutations, even for rare, relatively homogeneous Mendelian diseases such as DBA. While we were able to carefully calibrate variant quality between gnomAD and our cohort, joint calling of genotypes would have likely improved our power, as we erred on the side of being more conservative in our calibrations. Furthermore, as our ability to discriminate between benign and pathogenic variants improves, so will our ability to identify causal genes in Mendelian diseases. Assuming sufficient ascertainment of causal genetic variation, including limited instances of somatic reversion in whole blood, our results suggest that there is no single remaining gene with mutations detectable by WES that explains a large fraction (>5%) of the remaining cases, as we would have almost certainly detected a burden of LoF or missense mutations in a gene of this character, given the cohort sample size. This leads us to believe that a large percentage of the remaining causal variants are RP gene CNVs, as previous studies have observed that 15%–20% of case subjects harbor these, whereas our study detected only 10% since we did not use a comprehensive CNV screening assay. We also believe that we are only scratching the surface in identifying cryptic splice and large effect “non-coding” mutations (e.g., promoter, 3’ UTR, etc.) in RP genes. Comprehensively assaying CNVs will likely increase the diagnosis rate from 78% to 83%–88% and combined WGS and RNA-seq on remaining cases could push the rate over 90%.

Overall, our results and other recent reports suggest that at least 19 and perhaps 26 or more RP genes are involved in DBA pathogenesis (Figure 1E). This is ~1/3 of the genes that comprise the human ribosome, and mechanistic work from our group and others has suggested that these mutations predominately reduce ribosome levels, leading to a selective reduction in the translation of key genes involved in erythroid lineage commitment during hematopoiesis. However, there are still many unanswered questions. For example, it remains unclear whether CECR1 mutations result in an unrelated phenocopy or whether CECR1 lies on the same causal pathway as other DBA mutations. It will also be interesting to examine to what extent other identified variants, such as the LoF mutations in MYSM1, may interface with the GATA1 pathway that is critical for erythropoiesis. Additionally, our work has built on other studies by demonstrating robust genotype-phenotype correlations, the detailed mechanisms of which remain to be elucidated. Finally, we caution that the causality or pathogenicity of any specific variant is not certain without direct experimental evidence of its effect on the DBA phenotype, even for seemingly clear-cut LoF variants.

Accession Numbers

Variant call files and other information for all individuals reported in this study are available at dbGaP under accession number phs000474.v3.p2. All RNA-seq data from individual LCLs are available at NCBI GEO under accession number GSE119954.

Supplemental Data

Supplemental Data include six figure and eight tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.10.027.

Acknowledgments

We owe a great deal of gratitude to the individuals and families that were part of the cohort described here, without whom this study would not have been possible. We thank members of the Sankaran laboratory for valuable discussions and comments on this work. This work was supported by the National Institutes of Health grants R01 DK103794 and R33 HL120791 (to V.G.S.), R01 HL107558 and K02 HL111156 (to H.T.G.), UM1 HG008900 and R01 HG009141 (to D.G.M.), as well as a DBA Foundation grant and New York Stem Cell Foundation award (to V.G.S.). L.D.C., P.-E.G., and M.-F.O. are supported by ANR (ANR 2015 AAP genérique CE12- 0001 - DBA Multigenes), and the EuroDBA project is funded by the ERA-NET programme E-RARE3 (ANR-15-RAR3-0007-04). J.C.U. is supported by NIH grant 2T32 GM007226-43. V.G.S. is New York Stem Cell Foundation-Robertson Investigator.

Declaration of Interests

The authors declare no competing interests.

Received: July 10, 2018
Accepted: October 29, 2018
Published: November 29, 2018

Web Resources

GATK, https://software.broadinstitute.org/gatk/
gnomAD Browser, http://gnomad.broadinstitute.org/
References


