Published online 14 March 2019

Nucleic Acids Research, 2019, Vol. 47, No. 10 ***e59***
doi: 10.1093/nar/gkz164

Splice-Break: exploiting an RNA-seq splice junction algorithm to discover mitochondrial DNA deletion breakpoints and analyses of psychiatric disorders

Brooke E. Hjelm[1,2,*], Brandi Rollins[1], Ling Morgan[1], Adolfo Sequeira[1], Firoza Mamdani[1], Filipe Pereira[3], Joana Damas[4], Michelle G. Webb[2], Matthieu D. Weber[1], Alan F. Schatzberg[5], Jack D. Barchas[6], Francis S. Lee[6], Huda Akil[7], Stanley J. Watson[7], Richard M. Myers[8], Elizabeth C. Chao[9], Virginia Kimonis[9], Peter M. Thompson[10], William E. Bunney[1] and Marquis P. Vawter[1,*]

[1]Department of Psychiatry and Human Behavior, University of California-Irvine (UCI), Irvine, CA 92697, USA, [2]Department of Translational Genomics, Keck School of Medicine of USC, University of Southern California (USC), Los Angeles, CA 90033, USA, [3]Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Matosinhos 4050-123, Portugal, [4]The Genome Center, University of California-Davis, Davis, CA 95616, USA, [5]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA, [6]Department of Psychiatry, Weill Cornell Medical College at Cornell University, New York, NY 10065, USA, [7]The Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, MI 48109, USA, [8]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA, [9]Division of Genetics and Genomic Medicine, Department of Pediatrics, UCI, Irvine, CA, USA and [10]Southwest Brain Bank, Department of Psychiatry, Texas Tech University Health Sciences Center (TTUHSC), El Paso, TX 79905, USA

Received January 29, 2019; Editorial Decision February 25, 2019; Accepted February 28, 2019

ABSTRACT

Deletions in the 16.6 kb mitochondrial genome have been implicated in numerous disorders that often display muscular and/or neurological symptoms due to the high-energy demands of these tissues. We describe a catalogue of 4489 putative mitochondrial DNA (mtDNA) deletions, including their frequency and relative read rate, using a combinatorial approach of mitochondria-targeted PCR, next-generation sequencing, bioinformatics, post-hoc filtering, annotation, and validation steps. Our bioinformatics pipeline uses MapSplice, an RNA-seq splice junction detection algorithm, to detect and quantify mtDNA deletion breakpoints rather than mRNA splices. Analyses of 93 samples from postmortem brain and blood found (i) the 4977 bp 'common deletion' was neither the most frequent deletion nor the most abundant; (ii) brain contained significantly more deletions than blood; (iii) many high frequency deletions were previously reported in MitoBreak, suggesting they are present at low levels in metabolically active tissues and are not exclusive to individuals with diagnosed mitochondrial pathologies; (iv) many individual deletions (and cumulative metrics) had significant and positive correlations with age and (v) the highest deletion burdens were observed in major depressive disorder brain, at levels greater than Kearns–Sayre Syndrome muscle. Collectively, these data suggest the Splice-Break pipeline can detect and quantify mtDNA deletions at a high level of resolution.

INTRODUCTION

Large mitochondrial DNA (mtDNA) deletions causing human disease were first reported for 'mitochondrial myopathies' and Kearns–Sayre syndrome (KSS) roughly 30 years ago (1–3). Additional disorders that are now (at least partially) attributed to large mtDNA deletions include Pearson syndrome (PS), chronic progressive external ophthalmoplegia (CPEO), Leigh syndrome, and diabetes mellitus (4–11). Even in these 'hallmark disorders' of mtDNA deletions, there are currently no perfect genotype-phenotype correlations between a specific deletion and a particular symptom (or its age of onset) because these phenotypes are highly influenced by deletion abundance (het-

*To whom correspondence should be addressed. Tel: +1 323 442 7799; Email: bhjelm@usc.edu
Correspondence may also be addressed to Marquis P. Vawter. Tel: +1 949 824 9014; Email: mvawter@uci.edu

eroplasmy rate) and the type of tissue(s) affected, which can vary dramatically between subjects (6–13). Heteroplasmy rate, deletion size, and relative deletion position, however, have been correlated with measures of disease severity and age of onset (13). The clinical complexity observed in mitochondrial pathologies has prompted a number of investigations into whether mtDNA aberrations (including deletions) may be responsible for other human disorders, especially in diseases that present with neurological or muscular dysfunction, are progressive or degenerative, and/or affect multiple tissues and organ systems (9–12). To date, >800 human mtDNA deletions have been described in the scientific literature and have been curated in the MitoBreak online database (14).

Mitochondria have the principal responsibility of generating ATP by oxidative phosphorylation to provide every nucleated cell in the body with the energy it needs to function properly. Mitochondria contain their own DNA, a 16,659bp double-stranded circular genome that includes 13 protein-coding genes that make up the respiratory chain complexes (I–V), along with 22 transfer RNAs and 2 ribosomal RNAs essential for the translation of these important polypeptides (9–12). Although the mitochondrial genome is relatively small and contains few genes compared to any nuclear chromosome, its genomic evaluation is complicated by the fact that mtDNA is a polyploid feature of eukaryotic cells, with up to several thousand copies of the mitochondrial genome per cell (9,10). Under normal conditions, this polyploidy provides an excellent backup system for ensuring the cellular energy demands of a particular tissue are met because even if a number of mitochondrial 'energy generators' are defective, the remaining pool of functional mitochondria will prevent a total blackout. Thus large mtDNA deletions will only lead to a disease phenotype if they supersede a 'threshold' where the proportion of defective mitochondria cannot be functionally rescued by the remaining healthy mitochondria (8–12).

Mitochondrial functions are under the genetic jurisdiction of both the nuclear genome and the mitochondrial genome. Interestingly, defects in at least 18 nuclear genes (*POLG, POLG2, TWNK, RNASEH1, MGME1, DNA2, TK2, DGUOK, RRM2B, TYMP, SLC25A4, MPV17, OPA1*, *MFN2, C10orf2, SAMHD1, SPG7* and *AFG3L2*) have been associated with mtDNA deletions, suggesting these nuclear-encoded proteins affect the mtDNA replication and repair machinery, nucleotide pool, or fusion processes in such a way that the mitochondrial genome becomes more susceptible to deletion formation and/or accumulation (15–21). Single nucleotide variants (SNVs) in mtDNA and structural variants such as large deletions can be either maternally inherited or may occur *de novo* (9,10,12,22). While pathogenic SNVs have been observed in both homoplasmic and heteroplasmic states, large deletions in the mitochondrial genome are exclusively heteroplasmic (9,11,12). Thus, the historical evaluation of large mtDNA deletions in patient samples/affected tissues has largely focused on relative heteroplasmy rate and breakpoint identification.

The traditional methods utilized to assess mtDNA deletion heteroplasmy rate include southern blot analysis and quantitative PCR (qPCR). In the case of southern blotting, two conditions must be met: (i) enough DNA can be directly isolated from the sample (PCR amplification is biased towards smaller products) and (ii) the deletion must be present at a high enough rate for it to be detected (23–25). The accumulation of a specific mtDNA deletion over time (i.e. clonal expansion) may occur during development and affect an entire tissue, or may occur within single cells, so meeting these two conditions simultaneously may not always be possible. Alternatively, qPCR has been used to assess heteroplasmy rates using several strategies. Previously identified mtDNA deletions have been quantified and compared to the wild type DNA molecules using primers specifically designed to target these molecules. However, this strategy requires an *a priori* hypothesis about a specific deletion and its corresponding set of breakpoints. The 4977 bp 'common deletion', for example, has been detected in muscle, brain, skin, etc. as an effect of aging, UV exposure, doxorubicin treatment, and in association with a number of disorders (KSS, CPEO, patient's with *POLG* mutations, Parkinson's Disease, Alzheimer's Disease, Huntington's Disease, diabetes mellitus; see Human 'common deletion' page on MitoBreak website http://mitobreak.portugene.com for full list of references) (14). Detection of the 'common deletion' in so many disorders, however, is more likely the result of detection bias as this deletion is well known and easily quantifiable by qPCR. In addition, several groups have utilized methods that target the *ND1* and *ND4* genes by qPCR to determine their relative abundance, based on the assumption that the former is rarely deleted and the latter is commonly deleted across a wide-spectrum of mtDNA deletion breakpoints, and that this relative ratio may be used as a generic representation of deletion heteroplasmy rate (24–27). There are, however, large mtDNA deletions that encompass the ND1 gene so using this gene as a control region may not be the best approach for all cases (14,28).

Breakpoint identification has largely relied on Sanger sequencing, often using a PCR product that encompasses an mtDNA deletion breakpoint. PCR amplification and Sanger sequencing still remains an efficient strategy for identifying breakpoints when a single deletion is enriched throughout a patient's cells/tissues. However, there are many cases where multiple species of deletions (i.e. a variety of unknown breakpoints) would be hypothesized to occur in a tissue and thus would require a laborious workflow that includes plasmid cloning and selection (and/or single-cell isolation) in order to isolate and enrich each molecule sufficiently for Sanger sequencing (23,26). There have also been a few methods reported that use next-generation sequencing (NGS) data to identify mtDNA deletions; however, we believe these methods either lack vigor in regard to deletion quantification and breakpoint identification (29,30), or have been insufficiently tested with real mtDNA breakpoints and their associated repeat sequences (31). We have compared our bioinformatics pipeline to the MitoDel tool (31) using both DNA and RNA alignment algorithms in order to assess the effects of read mapping strategy and downstream filtering approaches on mtDNA deletion detection. Additional computational methods we have identified that may warrant future comparisons with our pipeline include a customized Perl script provided by Zambelli *et al.* (32), and

the eKLIPse tool recently published by Goudenège *et al.* (33).

The technical limitations described for both relative quantification of mtDNA deletions and the identification of deletion breakpoints motivated us to develop a pipeline that would generate this information in a high-throughput, high-resolution manner. To accomplish this, we developed a combinatorial approach of mitochondria-targeted PCR, next-generation sequencing (NGS), existent bioinformatics tools, post-hoc filtering, annotation, and validation steps. Our bioinformatics pipeline uses MapSplice, an RNA-seq splice junction detection algorithm, which we have exploited to detect mtDNA deletion breakpoints rather than mRNA splices (34). For simplicity, we refer to the entire NGS and bioinformatics pipeline as Splice-Break, and describe the methodology in detail in this paper. In addition to assessing artificial data for sensitivity and specificity, we also present the results of the first human study using this pipeline where we evaluate 93 human samples of post-mortem brain and blood from subjects with and without psychiatric disorders and compare these results to 'hallmark disorders' of mtDNA deletions, specifically KSS muscle and PS blood.

MATERIALS AND METHODS

Subjects

This study included analyses of 93 samples obtained from the Southwest Brain Bank (SBB) from the Department of Psychiatry at Texas Tech University Health Sciences Center and the University of California-Irvine (UCI) Pritzker Brain Bank at the UCI School of Medicine. A summary of the subjects' sex, age, diagnosis (psychiatric disorder) and the brain regions analyzed can be found in Supplemental Table S1, and in the Gene Expression Omnibus (GEO): GEO accession GSE118615. Detailed extraction procedures can be found in the online Supplemental Methods. In addition, DNA from two subjects with 'hallmark disorders' of mtDNA deletions, specifically KSS muscle and PS blood, were obtained from Dr Virginia Kimonis and Dr Elizabeth C. Chao; these subjects were previously determined to have large mtDNA deletions via a clinical diagnostic pipeline that utilized Southern Blot analysis (data not shown).

Long-range mitochondrial PCR

Prior to NGS library preparation, the mtDNA was enriched for each sample using a long-range (LR) PCR that utilizes back-to-back primers that hybridize to the control region of the mitochondrial genome (35). Primer sequences used were 5′-CCGCACAAGAGTGCTACTCTCCTC-3′ and 5′-GATATTGATTTCACGGAGGATGGTG-3′ (Integrated DNA Technologies, Coralville, IA, USA), for the forward and reverse primers, respectively. Each sample was enriched for mtDNA in a 50 μl, 30-cycle LR PCR reaction; 5 μl of each PCR product was subsequently used for agarose gel electrophoresis to confirm the PCR was successful (Supplemental Figure S2). Detailed procedures for PCR conditions and purification can be found in the online Supplemental Methods.

Library preparation and sequencing

Libraries were prepared using the TruSeq Nano DNA HT Library Preparation Kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Libraries were sequenced as 150-mer paired-end reads on a nonpatterned flowcell using the Illumina HiSeq 2500 in 'rapid mode'; the UCI Genomics High Throughput Facility performed both sequencing and demultiplexing of FASTQ files.

Bioinformatics

Pre-alignment. We removed PCR replicates and reads that mapped to the TruSeq adapter sequence prior to alignment using the BBTools suite (BBMap v.35.14) (https://jgi.doe.gov/data-and-tools/bbtools) (see Supplemental Methods for details).

Alignment and methods comparison. We used MapSplice (v.2.1.18), an RNA-Seq splice junction algorithm from Dr. Jinze Liu's group at the University of Kentucky (http://www.netlab.uky.edu/p/bioinfo/MapSplice2), for our mtDNA alignment process (34). Details about our rationale to use an RNA-Seq tool can be found in the online Supplemental Methods. We compared our process to an alternative mtDNA deletion detection method (MitoDel (31)) in addition to several alignment algorithms- MapSplice (34), TopHat (36), STAR (37) and BWA (38). The details of these comparisons can be found in the Results and online Supplemental Methods.

MapSplice. Our finalized Splice-Break pipeline used MapSplice (34) for alignment- a list of additional bioinformatics tools, command options, parameters and computational resources used can be found in the Supplemental Methods. Two output files from the MapSplice algorithm were used for further analysis: junctions.txt and alignments.bam. To facilitate the use of this pipeline, we provide a bash script called Splice-Break.sh (included in Supplemental File S3) that will perform automated filtering, normalization and annotation of mtDNA deletion breakpoints using these two, aforementioned MapSplice output files. A full description of our benchmark positions used for normalization, generation of coverage plots and normalization processing of 'junctions' can be found in the online Supplemental Methods.

Artificial data. We generated artificial FASTA reference files that correspond to five single mtDNA deletions detected in our un-filtered dataset, as well as a complex artificial FASTA reference file that contained 60 different deletions all mixed together. We used DWGSIM (v.0.1.10) (https://github.com/nh13/DWGSIM/wiki/Simulating-Reads-with-DWGSIM) to create artificial paired-end FASTQ files of 150-mer read lengths with precalculated proportions of wild type and deleted mitochondrial genomes while accounting for their differences in genome size. These artificial datasets were used for analysis of alignment method and options as well as for optimization of filtering steps; this analysis is described further in the online Supplemental Methods and Results.

Filtering. We employed three filtering steps to prune and prioritize our list of putative mtDNA deletion breakpoints (Supplemental Figure S4). The first step removed any deletion breakpoints that had a minimum read overhang length <20 bp. The second filtering step was to remove any putative deletion call where either the 5′ breakpoint or 3′ breakpoint fell within 500 bp of the 16 kb long-range PCR primer start positions (NC_012920.1 356-15926; modified NC_012920.1 16070-500). The third and last filtering step was to remove any putative deletion breakpoint that was detected in less than two independent samples (i.e., singleton calls), with the exception that singleton deletions discovered at a deletion read % of 5% or more would be retained pending they could be validated by Sanger sequencing. This last filtering step is not incorporated into the Splice-Break.sh script, but rather, was included here for a more conservative approach to creating a catalogue of mtDNA deletion breakpoints. Additional details are described in the online Supplemental Methods and in the README file that accompanies the Splice-Break.sh script (Supplemental File S3). Analysis scripts and code for the bioinformatics portion of Splice-Break are also located on GitHub: https://github.com/brookehjelm/Splice-Break/.

Quantitative PCR

Both the wild type mtDNA and the 4,977bp 'common deletion' mtDNA were evaluated by quantitative PCR (qPCR). Specifics regarding the qPCR conditions, methods and analysis can be found in the online Supplemental Methods.

Sanger sequencing validation

Deletions. The Sanger sequence flanking each (targeted) mtDNA deletion breakpoint, along with the primers used for amplification of the deleted molecule and the Sanger sequencing reactions are shown (Supplemental Figure S5). Once 'high frequency' or 'high impact' deletions were identified, Sanger sequencing was performed on both the 16 kb LR PCR product (that was used for library preparation) and the corresponding genomic DNA to validate the deletions. Specifics regarding primer design and PCR conditions can be found in the online Supplemental Methods.

Cumulative deletion metrics

We calculated three cumulative deletion metrics that can be used to investigate the pooled effect of all deletions. The first was the cumulative deletion read %, which is the summation of the read %'s determined for all deletions detected in that sample. The second was the # of deletions (per 10K coverage), which was calculated by counting the number of deletion species (unique sets of breakpoints) in that sample, then dividing by the benchmark coverage and multiplying by 10000×. The third cumulative deletion metric we determined was % burden per deletion, which was calculated by taking the cumulative deletion read % divided by the number of deletion species detected (raw data, not normalized). Although the deletion read %'s are already normalized to benchmark coverage, it is important to also include the benchmark coverage as a correction factor in any statistical tests as all three of these cumulative deletion metrics will be dependent on the depth in that sample regardless of normalization.

Statistical and graphical analysis

Details about all statistical and graphical analyses are provided in the online Supplemental Methods.

RESULTS AND DISCUSSION

Splice-Break pipeline

We developed a pipeline called Splice-Break that combines LR (∼16.6 kb) PCR amplification of the mitochondrial genome with NGS and existent bioinformatics tools to detect mtDNA deletion breakpoints (Figure 1). We utilized MapSplice, an RNA-seq splice junction detection algorithm, for sequence alignment and discovery of putative mtDNA deletion breakpoint 'junctions' (34). Our human subject cohorts included 93 samples (from 41 subjects) obtained from the SBB and UCI Brain Banks (Supplemental Table S1). For all 93 samples full-length mitochondrial wild type genomes were successfully amplified, as evidenced by a ∼16.6 kb DNA band following the LR PCR. Three brain samples (from two subjects) also displayed smaller DNA bands (∼2–10 kb) following the PCR, suggesting these samples may harbor large mtDNA deletions (Supplemental Figure S2). After MapSplice alignment of NGS reads (using libraries prepared from LR PCR products), we observed average benchmark coverage of 19899 ± 12301× (mean ± SD; range 2764–82596×) for all 93 human subject samples. Coverage plots of all human samples, along with the average benchmark coverage for each sample, are shown in Supplemental Figure S6. We have organized these plots by subject for an easy visual comparison of paired samples (i.e. multiple tissues or brain regions from the same donor), and have included the SBB and UCI Brain Bank ID's for each sample that correspond to those shown in GEO (GSE118615).

We processed the MapSplice junctions.txt files, containing the putative mtDNA deletion breakpoints, through a series of filtering steps that were devised to remove false-positive and low-confidence calls. Overall, these three filtering steps removed 90.32% of the initial junction calls, and resulted in an average of 346 and 61 unique mtDNA deletions for the postmortem brain and blood samples, respectively (Supplemental Figure S4). Our finalized catalogue contains 4489 putative mtDNA deletions, which is >5-fold more than all the human mtDNA deletion breakpoints currently described in the MitoBreak database (14). This catalogue is weighted to focus on mtDNA deletions that fall within genes, and only includes those with breakpoints between positions 357–15925 (NC_012920.1).

Artificial data

We generated a series of artificial FASTA reference sequences that reflected the predicted sequence of mtDNA molecules with large deletions, and combined deleted and wild type sequences at different ratios in order to determine method sensitivity and specificity (Figure 2). Analysis of the

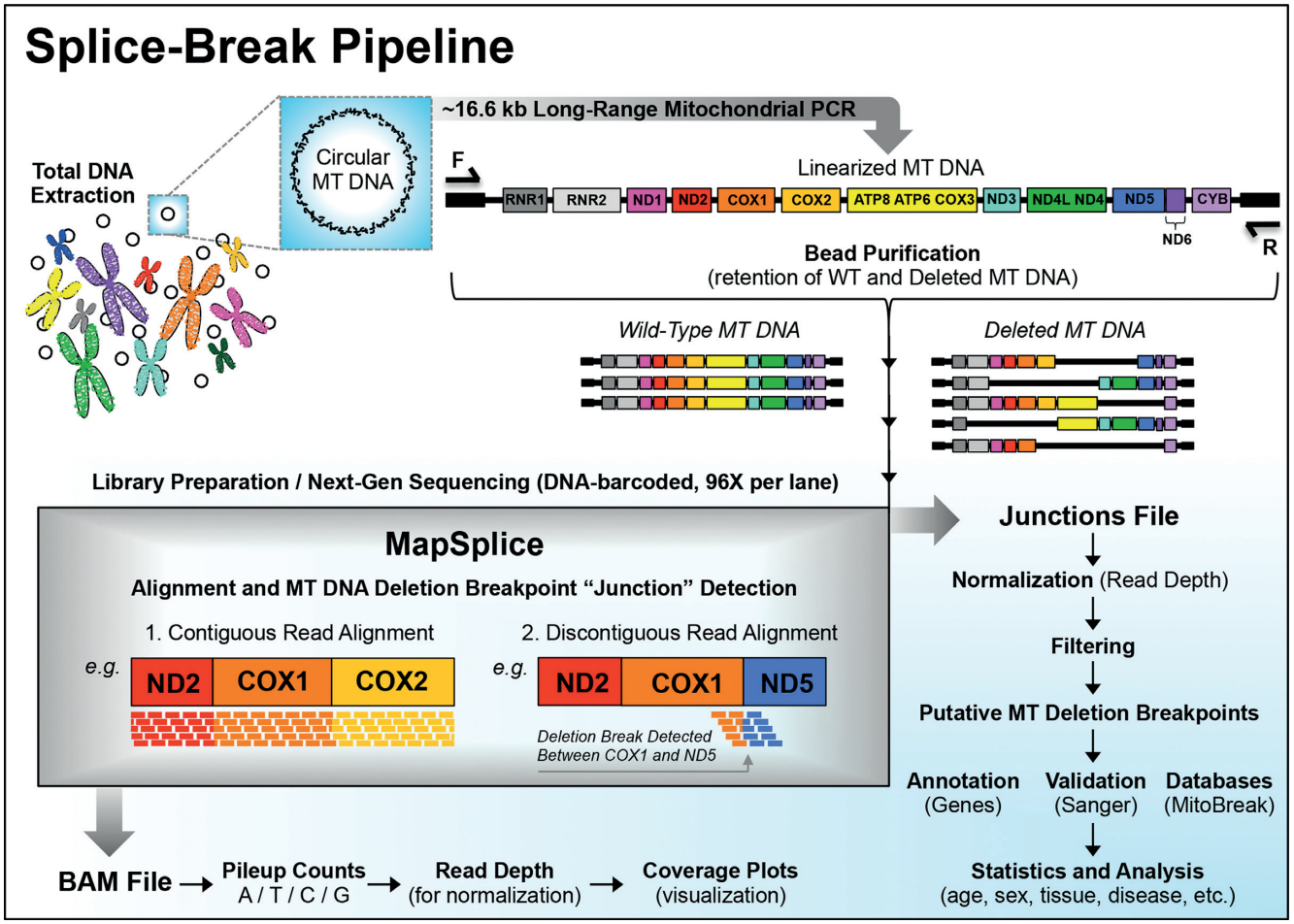

Figure 1. Summary of the Splice-Break pipeline. This pipeline integrates long-range (LR) PCR of the mitochondrial genome with bead purification, NGS library prep and multiplex sequencing to generate mitochondrial sequence data that contains mtDNA deletion breakpoints. Alignment and mtDNA deletion breakpoint 'junction' calling were accomplished using MapSplice, an RNA-seq splice junction detection algorithm (34).

4977 bp 'common deletion' (MapSplice breakpoints 8471–13449) demonstrated a drop in coverage between the deletion breakpoints was visible at a deletion rate of 5% or more, but deletions at a lower read rate (0.01–1%) are difficult (if not impossible) to detect by visual inspection of a coverage plot alone (Figure 2A). The average test sensitivity for the 'common deletion' was 70.25%, and displayed a high level of consistency ($R^2 = 0.99998$) across a wide range of deletion rates (0.1–50% tested). However, no 'common deletion' was detected when spiked-in at a heteroplasmy rate of 0.01%, which suggests deletions at or below this level might be missed entirely (Figure 2B). Deletions with MapSplice breakpoints of 6335–13999, 6329–13994, 7816–14707 and 1243–15340 were also evaluated. The coverage plots for all of these deletions reflected a sharp drop in coverage that corresponded to the breakpoints for that specific deletion (Figure 2C), and similar to results of the 'common deletion' there was a high level of consistency (R^2; range 0.99995–0.99998) in the test sensitivity for each deletion across a wide range of deletion rates (0.1%, 1%, 10% were tested). Interestingly, however, the test sensitivity was different for each of the five deletions tested (range 38.8–74.6%) (Figure 2D). We believe this variance in sensitivity is a reflection on how MapSplice handles reads that contain the various repeat sequences associated with mtDNA deletions, and that a fraction of reads (or read fragments) that actually span a deletion junction will incorrectly map to the wild type molecule due to an insufficient overhang length or capacity to match perfectly with the incorporation of a SNV or indel. In fact, we have observed false-positive, heteroplasmic SNV calls and indels around these repeat sequences in samples that contain certain deletions (data not shown); as such, we suggest that this pipeline be used only for analysis of large mtDNA deletions and not for SNVs or small indels. Likewise, this pipeline cannot be used for assessment of mtDNA copy number or mtDNA depletion pathologies because our LR PCR enrichment and NGS process (that includes normalization of libraries) limits our ability to assess nuclear DNA in an unbiased manner, which would be necessary for mitochondrial copy number determination.

We additionally tested the specificity of the MapSplice algorithm to call mtDNA deletions by examining the breakpoint 'junctions' and determining the proportion of reads that were called at the correct position. Prior to any filtering steps, all five deletions were detected at a high level of specificity (92.7–96.1%); however, for deletions present at a rate

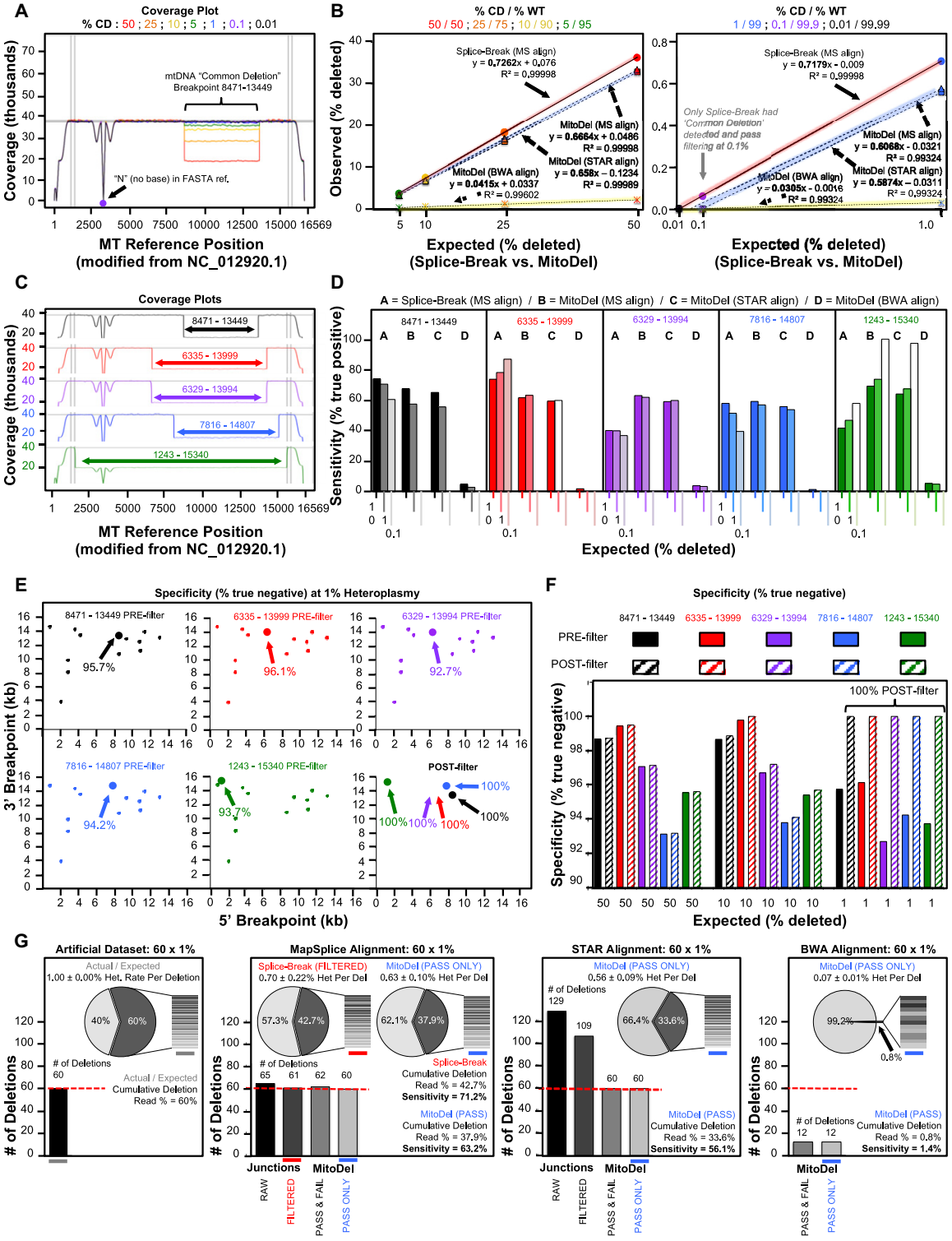

Figure 2. Analysis of artificial data. (**A**) Overlaying coverage plots of the mtDNA 'common deletion' from mixed ratios (0.01–50%) of deleted and wild type (undeleted) sequences, and (**B**) sensitivity analysis from this data and comparisons between Splice-Break (MapSplice (MS) alignment) and MitoDel (MS, BWA or STAR alignment). (**C**) Coverage plots of five Sanger-validated mtDNA deletions from 1:1 ratios of deleted and wild type (undeleted) sequences (50% each), and (**D**) sensitivity analysis for these deletions from a series of mixed ratios (1–50%) and comparisons between Splice-Break (MS alignment) and MitoDel (MS, BWA or STAR alignment). (**E**) Scatter plots of all junction calls observed for these five deletions before and after filtering steps when spiked-in at a heteroplasmy rate of 1%; arrows point to the actual breakpoint positions for that deletion. (**F**) Specificity (% true negative) assessed for these five deletions before and after filtering steps using a series of mixed ratios (1–50%). (**G**) Analysis of complex artificial deletion dataset combining 60 mtDNA deletions each at a rate of 1% (total expected heteroplasmy rate of 60%) and comparisons between Splice-Break (MS alignment) and MitoDel (MS, BWA or STAR alignment). Pie charts display detected sensitivity of cumulative deletion read %, and bar graphs display number of unique deletion breakpoints observed with each approach.

of 1% or less, the test specificity increased to 100% for all five deletions when deletion calls with a read overhang <20 bp were removed (Figure 2E and F). More than 99.7% of all deletions discovered in our biological cohort of postmortem brain (i.e. tissue homogenate) and blood samples were determined to have a deletion read rate <1%, so this analysis of artificial data suggests the vast majority of real mtDNA deletions would be called with perfect specificity using this method. We also evaluated wild type (non-deleted) consensus sequence generated from each the 41 subjects evaluated in this study, and likewise found that the filtering steps previously described removed all false-positive deletion calls that were generated in different mitochondrial haplogroup backgrounds (data not shown).

Finally, we generated a complex artificial FASTA file that contained 60 unique mtDNA deletions, each at a heteroplasmy rate of 1%, for a combined deletion rate of 60%. This complex artificial file is a better representation of what we expect to observe in frozen brain tissue samples where many deletion species are expected to be present. A list of the 60 mtDNA deletions combined in this complex artificial FASTA fie can be found in Supplemental Figure S7. Overall, we observed similar sensitivity to that of the 'common deletion', with a detection rate of $0.70 \pm 0.22\%$ for each deletion spiked in at 1%, and a cumulative deletion rate of 42.7% in this file with an expected heteroplasmy rate of 60% (sensitivity 71.2%) (Figure 2g).

Bioinformatic methods comparisons

To confirm that an RNA-Seq alignment tool must be able to detect non-canonical splice sites in order to be repurposed as an mtDNA deletion detection tool, we first compared MapSplice (34) to TopHat (36). TopHat junction calling is restricted to canonical splice sites while MapSplice is not (34,36,39). We evaluated six brain samples within a 1kb window for both the 5′ (8–9 kb) and 3′ (13–14 kb) breakpoints. This window was chosen for this comparison as it would theoretically harbor deletion calls for the 4977 bp mitochondrial 'common deletion', which we predicted would be the most likely to occur in many/all samples, and because additional mtDNA deletions have been previously identified in this region (14). Only MapSplice detected a deletion breakpoint (junction call) in all six samples tested that fell within the repeat regions of the 4977 bp 'common deletion' (Supplemental Figure S8).

We also tested a mtDNA deletion detection tool (MitoDel) in conjunction with two RNA-Seq alignment methods (MapSplice and STAR), as well as with the suggested DNA aligner (BWA), and compared these results to the sensitivities described above for our Splice-Break pipeline (31,34,36–38). Our assessment of the 'common deletion' demonstrated a robust effect of alignment method, with the poorest detection sensitivity (3–4%) resulting from MitoDel after BWA alignment (i.e. BWA mem with default settings) (Figure 2B). This is not particularly surprising given that DNA alignment methods often assume reads will align perfectly and contiguously across the genome with only small gaps being allowed for SNP and indel incorporation, and are not necessarily designed to handle large read splits of several kilobases like an RNA-Seq aligner must due to splicing. As such, MitoDel performed considerably better when used in conjunction with MapSplice or STAR; however, the detection sensitivity was still poorer for these approaches than it was for our Splice-Break pipeline that uses MapSplice (Figure 2B). Assessment of five different sets of mtDNA breakpoints also demonstrated that only Splice-Break was able to consistently call each deletion when present at a rate of 0.1% (Figure 2D). This effect is most likely due to the requirement of the MitoDel tool to observe a certain number of split reads in order for a deletion to pass its filtering process, which is not a requirement of Splice-Break; this makes the Splice-Break pipeline a particularly useful approach for assessing mtDNA deletion breakpoints in homogenate tissue where many deletions would be expected to occur in parallel at a low rate.

Lastly, assessment of our complex artificial FASTA file that contained 60 different deletions each at 1% heteroplasmy also demonstrated our Splice-Break bioinformatics process had the highest detection sensitivity when compared to MitoDel (with any alignment algorithm), and our attempts to use similar filtering approaches following STAR alignment were unsuccessful as many more deletion positions were detected than the 60 that were expected (i.e. calls within a repeat from a single mtDNA deletion would be annotated across multiple positions rather than collapsing to a single set of breakpoints) (Figure 2G). These results support our suggestion to use the Splice-Break pipeline for samples that may contain a large number of mtDNA deletions at low rates (e.g. homogenate brain/muscle tissue, previously extracted DNA, etc.), but also suggest that the MitoDel tool approach was sufficient to detect deletion breakpoints if (a) only 1 (or a small number) of deletions are expected to occur at a high heteroplasmy rate (e.g. 'single deletion' subjects with deletion pathologies, or DNA obtained after single-cell isolation) and (b) an RNA-Seq alignment tool is utilized. We also processed all 93 postmortem samples in this study with MitoDel following MapSplice alignment, in order to further compare our Splice-Break filtering process to MitoDel without the influence of alignment algorithm. Box-and-whisker plots of each sample's deletion %'s demonstrates that Splice-Break detected many deletions at a low rate (<1%), but that the deletions at higher heteroplasmy rates were detected by both approaches; this also had influence on the number of deletion species detected (Supplemental Figure S9). Later, we describe if and how this tool would have affected our biological conclusions of tissue, age, and paired samples, and which method correlates most closely to qPCR results of the 'common deletion'.

Exome comparisons

In addition to our assessment of bioinformatic methods, we also evaluated how mtDNA deletion discovery was influenced by our sequencing strategy. For this, we compared 10 dorsolateral prefrontal cortex (DLPFC) samples that were sequenced by exome sequencing to matched samples prepared with our pipeline that enriches the mtDNA molecules by LR PCR followed by NGS library preparation; both sequence files were processed through the same bioinformatics pipeline described for Splice-Break. We detected significantly more mtDNA deletions at a much higher rate with

our LR PCR approach than we did with exome capture (Supplemental Figure S10). Specifically, the most common mtDNA deletion that we identified (and Sanger validated) in our cohort (6335–13999) was detected in all 10 brain samples, with a range of 1–183 reads showing this deletion per sample (deletion read % range ∼0.01–1.75%) using the LR PCR approach. Contrarily, this deletion was only detected in one sample following exome capture, and this sample only had a single read supporting the deletion. Similar results between enrichment methods were also observed for our second and third most common mtDNA deletions (7816–14807 and 8471–13449 'common deletion') (Supplemental Figure S10). Moreover, our Splice-Break pipeline with LR PCR detected an average of 159 unique deletions per sample with an average cumulative deletion read rate of 21.35%, while exome capture detected only an average of 0.7 unique deletions per sample (i.e. 7 total deletions in 10 samples) with an average cumulative deletion read rate of 0.098% (Supplemental Figure S10). These results are perhaps not surprising given that the traditional probes used for exome capture will not, by design, encompass a deletion breakpoint and thus reads that flank these breakpoints will be less likely to compete with wild type molecules during the enrichment process. However, we do believe that LR PCR is not the only worthwhile approach to enriching the mitochondrial genome relative to the nuclear molecules, and other strategies aimed at selective digestion of the linear nuclear chromosomes or targeted isolation of the mitochondrial organelle may also provide cost-effective solutions. Moreover, exome capture strategies may be inefficient as is, given the current probe designs, but catalogues of mtDNA deletion breakpoints such as the one we describe here may be used as a template for designing additional probes and assays that will capture these split-read molecules.

Quantitative PCR comparisons

We compared our deletion read % obtained from Splice-Break (and following benchmark normalization) to heteroplasmy rates of the 4977 bp 'common deletion' obtained by qPCR across 61 samples (DLPFC and ACC pairs) in the SBB cohort (Supplemental Figure S11). We also compared both the qPCR results, and our Splice-Break values, to the deletion results obtained from the MitoDel tool following MapSplice alignment (Supplemental Figure S11). The qPCR results demonstrated these 61 brain samples had a 'common deletion' heteroplasmy rate of 0.12 ± 0.11% (mean ± SD; range 0.01–0.50%), while the Splice-Break results displayed a deletion read % of 0.40 ± 0.30% (mean ± SD; range 0.01–1.24%) and the MitoDel results displayed a deletion read % of 0.26 ± 0.26% (mean ± SD; range 0.00–0.93%). For this deletion specifically, our Splice-Break deletion % results were ∼4 ± 2-fold greater than the heteroplasmy rates obtained by qPCR, which we believe reflects an amplification bias toward smaller (deleted) molecules during the 16 kb LR PCR performed prior to NGS library preparation. Most importantly, however, is that we observed a very high correlation (Pearson's correlation: $r = 0.8384$, $P = 3.43e^{-17}$) between our Splice-Break pipeline (MapSplice) results and this gold standard qPCR assay (Supplemental Figure S11); we obtained a similar (albeit not as good) correlation between the MitoDel data and qPCR (Pearson's correlation: $r = 0.7976$, $P = 1.44e^{-14}$). This high level of reproducibility between bioinformatics approaches can also be observed in correlation analysis of Splice-Break versus MitoDel (with MapSplice alignment) in these biological samples (Pearson's correlation: $r = 0.9554$, $P = 5.86e^{-33}$) (Supplemental Figure S11). In addition, the $\log_2$ fold differences for the 'common deletion' between samples were largely retained using both Splice-Break and qPCR methods (Supplemental Figure S11). Collectively, these results demonstrate that our Splice-Break pipeline can quantify the *relative* levels of mtDNA deletions with high confidence; however, our deletion read % described should not be interpreted as heteroplasmy rate directly due to PCR amplification bias and deletion-specific differences in test sensitivity.

The ∼4500 putative mitochondrial DNA deletions

We detected 4489 putative mtDNA deletions that met our filtering criteria. As a reminder, this only includes those with breakpoints between positions 357–15925 (NC_012920.1) of the mitochondrial genome as we filtered out the majority of the control region in this analysis. The gene positions of all 4489 deletion breakpoints, the distribution of deletion sizes, and the redundancy in breakpoint usage are shown in Figure 3. This analysis is specific to the number and positions of all the putative mtDNA deletion breakpoints and is not weighted based on relative deletion read % (i.e. all deletions shown are considered equal in this view). 5′ breakpoints were identified in every protein-coding gene and both ribosomal RNA coding regions, with the highest number of 5′ breakpoint found in COX1 ($n = 665$), ND1 ($n = 657$), ND2 ($n = 639$), and the gene cluster of ATP8-ATP6-COX3 ($n = 551$) (Figure 3A and B). 3′ breakpoints were also identified in every protein-coding gene and both ribosomal RNAs, with the highest number of 3′ breakpoints found in ND5 ($n = 1418$), CYB ($n = 1243$), ND6 ($n = 936$) and ND4/ND4L ($n = 479$) (Figure 3A and C). The fact that genes with the most 5′ breakpoints are not the same genes with the most 3′ breakpoints reflects our observation that most of the deletions we detected were large, and multigenic in nature. A histogram plot of all 4489 putative deletions by size likewise demonstrated a preponderance of large deletions, with the greatest number of deletions ranging in size of 7–8 kb ($n = 513$) (Figure 3D). In contrast, only 127 deletions were identified within our smallest two bins, ranging from 50 bp (the minimal 'intron' size used during MapSplice alignment) to 1 kb in size. This histogram of deletion sizes displayed a semblance of a normal distribution; however, this distribution did not meet the criteria of normality when evaluated by visual inspection of *Q*–*Q* plots (data not shown) or statistical analysis by Shapiro–Wilk test ($P = 1.5e^{-20}$). Finally, our analysis of the 4489 putative deletions demonstrated a redundancy in the usage of 5′ and 3′ breakpoint positions (i.e. a given 5′ breakpoint is often used in conjunction with a number of 3′ breakpoints (and vice versa) to generate a spectra of unique deletions) (Figure 3E and F). The five most common 5′ breakpoints (and number of deletions they were utilized in include positions 5368 ($n = 56$), 5336 ($n = 48$), 4167 ($n = 48$), 591 ($n = 47$) and

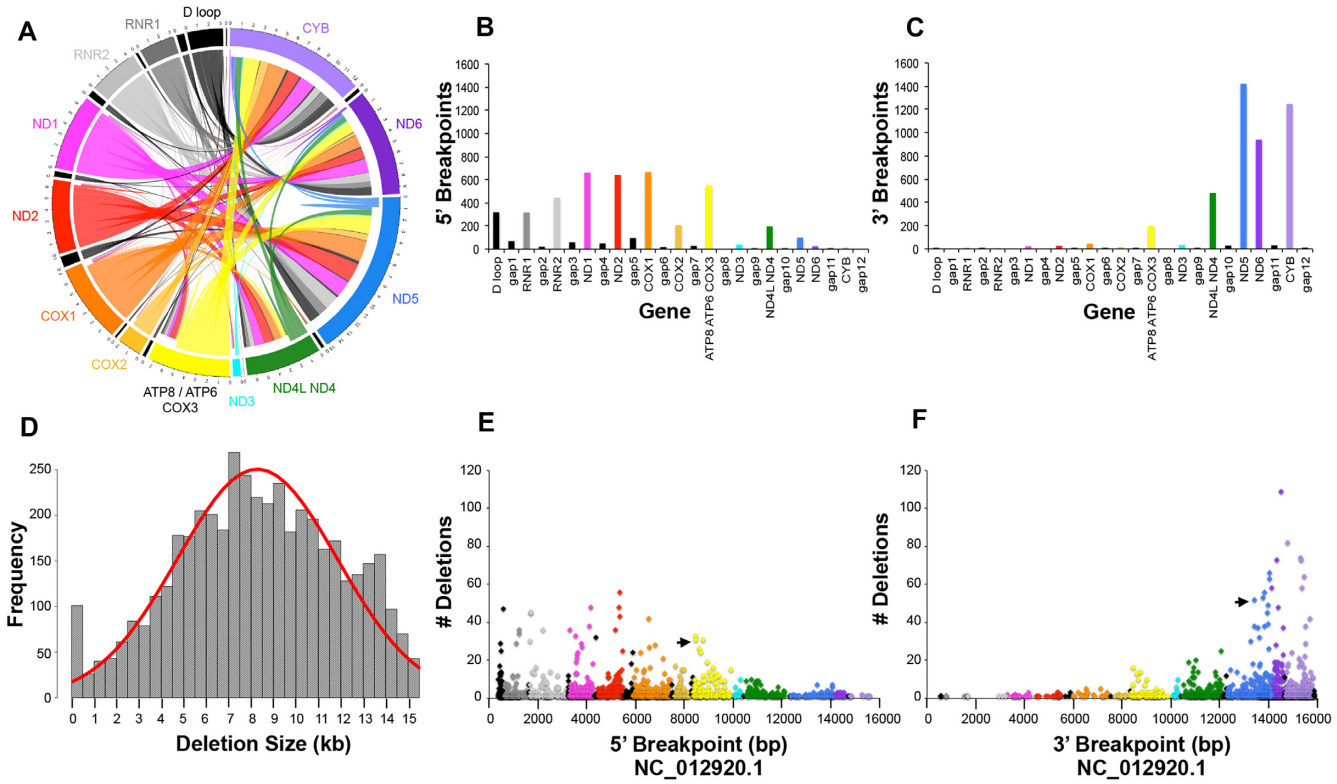

Figure 3. Global analysis of 4489 mtDNA deletions. (**A**) Chord diagrams display gene involvement across 4489 mtDNA deletions. Gene regions shown are not to scale with mitochondrial coordinates; the size of the gene shown represents the total number of deletion events (5′ breakpoints + 3′ breakpoints) within that gene. Ribbons are colored based on which gene contained the 5′ breakpoint. Gene locations for all (**B**) 5′ breakpoints and (**C**) 3′ breakpoints used in the 4489 mtDNA deletions; regions between protein-coding genes (that may contain several tRNAs) are referred to as gap1, gap2, etc. for simplicity. (**D**) Histogram of deletion sizes for all 4489 mtDNA deletions. Redundant usage of (**E**) 5′ and (**F**) 3′ breakpoints in multiple deletion species; arrows point to the 5′ and 3′ breakpoints for the 4977 bp 'common deletion', which were used in 31 and 52 deletions, respectively.

1698 ($n = 45$). The five most common 3′ breakpoints (and number of deletions they were utilized in) include positions 14529 ($n = 109$), 14807 ($n = 82$), 15335 ($n = 74$), 15382 ($n = 73$) and 14377 ($n = 73$). In addition to this global analysis (Figure 3), we also provide a supplemental table (Supplemental Data S12) that includes the 5′ and 3′ breakpoint positions of all 4489 mtDNA deletions, the deletion sizes, biological sample frequency, and several metrics of deletion read % (i.e. mean, median, standard deviation, minimum and maximum levels).

The top 30 most frequent deletions

We ranked our list of putative mtDNA deletions by the number of biological samples the deletion was detected in, and performed Sanger sequencing to validate the breakpoints for the 30 most frequent deletions (Supplemental Figure S5). All 30 of these deletions had breakpoints associated with perfect repeat sequences, often part of a larger, imperfect repeat (Table 1). We attempted to Sanger validate these deletions using both the bead-purified 16 kb LR mitochondrial PCR product and non-amplified, total genomic DNA; all 30 deletions were successfully validated by Sanger sequencing using the PCR product, and 14/30 of these deletions were additionally validated using the genomic DNA as well. The breakpoints identified by the MapSplice algorithm often fall within the direct repeat sequence, while the annotation of mtDNA deletion breakpoints is most commonly positioned to the last base of the 5′ repeat and the first distinguishable (non-repeat) base at the 3′ position; as such, we also provide the adjusted positions for these 30 deletions, which were used for overlap analysis and deletion submission in the MitoBreak database (14). Interestingly, 12/30 of the most frequent deletions were previously described in MitoBreak, with clinical features ranging from KSS, CPEO, *POLG* mutations, multisystemic mitochondrial disorders, aged tissues, and more (Table 1). This data suggests these common deletions are present at low levels in metabolically active tissues and are not exclusive to individuals with diagnosed mitochondrial pathologies.

Although these deletions were detected in a large proportion of our biological (mostly brain) samples, they were most often still detected at very low read rates (Supplemental Data S12). The most frequent deletion (MapSplice 6335–13999; adjusted 6341–14005) was detected in 92/93 biological samples at a deletion read % of 0.65 ± 0.57% (mean ± SD; range 0.002–2.73%). For reference, a deletion read % of 0.65% correlates with a mean detection of 130 reads for a sample with a benchmark coverage of 20000×. The 4977 bp 'common deletion' (MapSplice 8471–13449; adjusted 8482–13460) was detected in 85/93 biological samples at a deletion read % of 0.37 ± 0.32% (mean ± SD; range 0.011–1.47%). Many of the other 30 most frequent dele-

Table 1. The 30 most frequent mtDNA deletions. The position (5′-3′ breakpoints and gene), repeat sequences, sample frequencies, Sanger sequencing validation results, and overlap analysis with the MitoBreak database are shown (14). Deletions are sorted based on the number of samples they were detected in. 'MapSplice Breakpoints' displays the breakpoint positions called by the MapSplice algorithm, while 'Position adj.' shows the adjusted breakpoints based on how mtDNA deletions are typically annotated with regards to the repeat sequence. The bases called as MapSplice breakpoints are shown in red. Sanger sequence flanking each breakpoint and primers used for deletion validation are shown in Supplemental Figure S5

	Mitochondrial Deletion	Repeats and Breakpoints	No. Samples			Sanger Validation		MitoBreak Database		
No.	MapSplice Breakpoints 5' - 3'	imperfect repeats shown perfect repeat MapSplice break	Total n=93	Brain n=84	Blood n=9	PCR prod yes / no	gDNA yes / no	Database yes / no	Position (adj.) 5' - 3'	Diseases and/or Clinical Features
1	6335 - 13999 (COX1 - ND5)	CCTCCGTAGACCTAACC CCTCC TAGACCTAACC	**92**	84	8	yes	yes	yes	6341 - 14005	KSS; MNGIE; ad-PEO; Ptosis; Proximal limb weakness; encephalopathy; tubulopathy
2	7816 - 14807 (COX2 - CYB)	TCATCGCCCTCCCATCCC TCATCGACCTCCCCACCC	**86**	84	2	yes	yes	yes	7814 - 14805	Diffuse leukodystrophy; Aged tissues
3	8471 - 13449[a] *"The Common Deletion"* (ATP8 - ND5)	ACCTCCCTCACCA ACCTCCCTCACCA	**85**	83	2	yes	yes	yes	8482 - 13460	KSS; PEO; PS; ad/ar-PEO; ad/ar-PEO (POLGmut); Aged tissues; Unfertilized oocytes; Spermatozoa; Postmenopausal ovaries; Embryos; Cumullus cells; Controls; Parkinson Disease; Inclusion body myopathy (ad); Inclusion body myositis; Thyroid, Hepatic, Warthin's, Goiters tumors; Sporadic breast cancer and benign breast diseases; Reye-like; Alzheimer Disease; Huntington's Disease; Adrenal failure; Hemodialysis patients; Cyclic vomiting syndrome; Hearts exposed to doxorubicin; Pancytopenia; Microvesicular stetosis; Presbiacusis; Cerebral folate deficiency; Atrial fibrillation; Ataxia telangiectasia; Sporadic Amyotrophic Lateral Sclerosis; Toni-Debré-Fanconi syndrome; Chronic Kidney Disease; Polypodium treated patients; Ptosis, ophthalmoparesis; facial paresis; ophthalmoplesia; retinopathy; ataxia cerebellum syndrome; myopathy; tetraparesis; Skeletal Muscle Symptoms; multiorgan involvement; multiple deletion patient; Dilated cardiomyopathy; Chronic Fatigue Syndrome; Sensorineural Hearing Loss; MELAS; MERRF; Diabetes mellitus and deafness; Mitochondrial encephalomyopathy; Mitochondrial myopathy; COX deficiency; LS; Isolated mitochondrial myopathy; Sun-exposed skin
4	6545 - 13846 (COX1 - ND5)	ACCTCAAC ACC ACCTCAACTACC	**83**	82	1	yes	yes	yes	6551 - 13852	Multiple deletion patient (POLG1mut); Aged tissues; PEO, Parkinson
5	7126 - 14004 (COX1 - ND5)	CCTAGACCAAACCT CCTAGACCTAACCT	**83**	83	0	yes	yes	yes	7128 - 14006	Parkinson's Disease
6	7720 - 15821 (COX1 - ND5)	ACTCACAACAAAACTAA CTTCACAACAATCCTAA	**83**	83	0	yes	yes	no	7730 - 15831	
7	1105 - 13846 (RNR1 - ND5)	AGCCCTAAACCTCAAC AGCCCTAGACCTCAAC	**81**	76	5	yes	no	no	1102 - 13843	
8	6329 - 13994 (COX1 - ND5)	CCTCCGTAGACCTAACC CCTCC TAGACCTAACC	**78**	77	1	yes	yes	yes	6329 - 13994	KSS; PS; PEO; Aged tissues; Reye-like; MELAS; LS
9	7981 - 15503 (COX2 - CYB)	AGGCGACC AGGCGACC	**77**	75	2	yes	no	yes	7982 - 15504	KSS; PEO; Mitochondrial myopathy; MELAS; Ptosis; Proximal limb weakness; Chronic fatigue syndrome; Ataxia; Seizures; other neurological symptoms
10	983 - 13803 (RNR1 - ND5)	CTAAAACTCACC CTAAAACTCACA	**75**	72	3	yes	no	no	983 - 13803	
11	7499 - 14426 (TRNS1 - ND6)	CAACCCCATGGCCTCCATGACT CAACCCC TGACCCCCATGCCT	**74**	74	0	yes	no	yes	7506 - 14433	Multisystemic mitochondrial disorders
12	8471 - 14377 (ATP8 - ND6)	ACCA CCTACCTCCCTC ACCAATCCTACCTCCATC	**73**	72	1	yes	yes	no	8465 - 14371	
13	1714 - 15517 (RNR2 - CYB)	ACCTTAGCCAAACC ACCCTAGCCAA CC	**72**	67	5	yes	yes	no	1723 - 15526	
14	7863 - 15382 (COX2 - CYB)	ATC CCTCCC TTACC ATCACCTCCCATT CC	**72**	72	0	yes	no	no	7868 - 15387	
15	7126 - 14529 (COX1 - ND6)	AACCTACGCCAAAA AACCTCCCCCAAAA	**70**	69	1	yes	yes	no	7128 - 14531	
16	8574 - 13999 (ATP6 - ND5)	TCCTAGGCCTACCC TCCTAGACCTAACC	**70**	70	0	yes	no	no	8577 - 14002	
17	8624 - 14815 (ATP6 - CYB)	TCCCCACCTCCA TCCCCACC CCA	**70**	70	0	yes	no	no	8625 - 14816	
18	1127 - 13868 (RNR1 - ND5)	AAT CAACAAAACT AACCAACAAA CT	**69**	66	3	yes	no	no	1126 - 13868	
19	5368 - 14055 (ND2 - ND5)	CTCCACCTCAATCA CTCCACCTCCATCA	**69**	69	0	yes	no	no	5371 - 14058	
20	6219 - 13449 (COX1 - ND5)	ACCTCCCTC ACCTCCCTC	**69**	68	1	yes	yes	yes	6226 - 13456	KSS
21	7807 - 13999 (COX2 - ND5)	TC ATCCTAGTCCT TCCTCCTAGACCT	**69**	69	0	yes	yes	no	7805 - 13997	
22	8471 - 13584 (ATP8 - ND5)	CTACCTCCCT CTACCTCCCT	**69**	69	0	yes	yes	yes	8477 - 13590	Aged tissues; Multisystemic mitochondrial disorders
23	8624 - 15335 (ATP6 - CYB)	CCACCTCC CCACCTCC	**69**	69	0	yes	no	no	8628 - 15339	
24	8624 - 14055 (ATP6 - ND5)	ATCCCCACCTCCA ATCTCCACCTCCA	**68**	68	0	yes	no	no	8629 - 14060	
25	5368 - 15335 (ND2 - CYB)	ACTCCACCTC ACTCCACCTC	**67**	66	1	yes	no	no	5371 - 15338	
26	6545 - 14419 (COX1 - ND6)	AGACCGCAACC AGACCTCAACC	**67**	66	1	yes	no	no	6546 - 14420	
27	6840 - 11136 (COX1 - ND4)	GCTATCCCCACC CTTATCCCCACC	**67**	67	0	yes	yes	no	6851 - 11147	
28	7863 - 13449 (COX2 - ND5)	CCTCCCTTACCAT CCTCCCTCACCAT	**67**	67	0	yes	no	yes	7869 - 13455	KSS
29	7816 - 15382 (COX2 - CYB)	CCTCCCATCCC CCTCCCATTCC	**66**	66	0	yes	no	yes	7823 - 15389	PEO
30	3789-14807 (ND1 - CYB)	ACCTCTCCACCC ACCTCCCCACCC	**65**	64	1	yes	no	no	3792 - 14810	

tions described here were detected at a deletion read % even lower than these two deletions (Supplemental Data S12). These data suggest the majority of our biological samples harbor many of the same mtDNA deletions, but each deletion is present at varying, unpredictable (and usually low) levels—these observations further support our motivation to perform cumulative analyses that factor in all detected deletions and/or their additive deletion read %'s (described later).

The gene positions of the breakpoints associated with these 30 frequent deletions, the distribution of deletion sizes, the proportion of each base (A,T,G,C) used in the associated imperfect repeats, and the correlations between all these deletions, age, pH and PMI are shown in Figure 4. The 5′ breakpoints of these 30 deletions were identified in the first 6 protein-coding genes of the mitochondrial genome sequence (ND1, ND2, COX1, COX2, ATP8 and ATP6), one transfer RNA (TRNS1) that resides between COX1 and COX2, and both ribosomal RNAs (RNR1 and RNR2), with the highest number of 5′ breakpoints found in COX1 ($n = 9$), followed by COX2 ($n = 6$) (Figure 4A and D). The 3′ breakpoints of these 30 deletions were identified in the last four protein-coding genes of the mitochondrial genome sequence (ND4, ND5, ND6 and CYB) with the highest number of 3′ breakpoints found in ND5 ($n = 16$), followed by CYB ($n = 9$) (Figure 4A and D). We provide an artificial bed file (Supplemental Data S13) with the adjusted breakpoints of these 30 deletions for easy visualization of their genomic coordinates using a tool like the Integrative Genomics Viewer (IGV) (40) (Figure 4D).

A histogram plot of the 30 most frequent deletions by size shows that all of these deletions were large, ranging from ∼4–14 kb in size (Figure 4B). We additionally evaluated the imperfect repeat sequences (shown in Table 1) for these 30 deletions and compared the prevalence of each base in these repeats to the proportion used in the entire mitochondrial genome (Figure 4C). We detected a significant difference in proportion in all four bases, and show the base proportions and statistics for both the light (coding) and heavy strands of the mitochondrial genome (Figure 4C). Specifically in regards to the heavy strand, there was a decrease in the proportion of A, T and C bases within these imperfect repeats compared to what would be expected from the whole mitochondrial genome sequence, with the most abundant and significant decrease ($P < 1e^{-5}$) being observed from a loss of C bases within the imperfect repeats. These results were inversely related to the significant increase in the proportion of G bases ($P < 1e^{-16}$) within these imperfect repeats on the heavy strand. This significant enrichment of guanine (G) bases in the repeat sequences of our 30 most frequent deletions detected supports previous observations that mtDNA deletion breakpoints are associated with sequences with G-quadruplex forming potential (QFP) (41).

Lastly, we performed a Pearson's correlation test to evaluate the relationship between these 30 deletions, and pH, PMI, and donor age (Figure 4E and F). Only data from our DLPFC samples were used for this analysis ($n = 41$). Both a matrix of these correlation levels (Figure 4E) and a network plot of the significant correlations (Figure 4F) suggested the deletions we analyzed could be separated into two groups. The majority (23/30) of these frequent deletions had significant and positive correlations with one another (with few exceptions) and 18 of these 23 deletions also had a significant and positive correlation with age. The minority group (7/30) of these deletions also displayed a significant and positive correlation with one another (with few exceptions), but only 1 of these 7 deletions had a significant and positive correlation with age (Figure 4E and F). Upon examination of the 5′ and 3′ breakpoints of these two groups of deletions, we discovered that the majority group (23/30; 76.7%) of deletions would have retained the lagging strand origin-of-replication (OL) in the deleted molecule, while the minority group (7/30; 23.3%) of deletions that mostly did not correlate with age had the OL removed (Figure 4F). This latter result supports a previously published observation that ∼21.9% of mtDNA deletions detected in healthy or aged tissues have the OL removed (7). In addition, we observed almost no correlations with pH and PMI, with the exception that there was a negative correlation between them (as would be expected) and one deletion (MapSplice 1105–13846, adjusted 1102–13843) had a negative correlation with PMI suggesting this deleted molecule may be more susceptible to degradation. We observed no significant correlations between any of the top 30 deletions and the pH measures taken from the cerebellum of the same subject (Figure 4E and F).

High impact deletions

In addition to the top 30 most frequent deletions, we also characterized all of the 'high impact' deletions that occurred at a deletion read rate of 5% or more, which suggests the deletion was clonally expanded and may have a developmental origin. The 13 'high impact' deletions we describe were only detected at high levels in postmortem brain samples, and not in blood, and were identified in 14 brain tissue samples from 9 subjects with various psychiatric diagnoses (Table 2). These deletions were almost exclusively associated with perfect (and imperfect) repeat sequences, with the exception of one deletion (MapSplice 2973–15573) that had no identifiable repeat sequence flanking the breakpoint. Also interesting, one deletion (MapSplice 1148–15607) incorporated two tandem copies of a 6-mer repeat (CTACGA), which suggests neither the 5′ nor 3′ repeat was completely lost during deletion formation (Table 2). All 13 of these deletions were successfully validated by Sanger sequencing, including the two deletions mentioned above that did not have 'traditional' repeat breakpoints (Supplemental Figure S5). Together, these 13 'high impact' deletions were composed of both rare deletions (that were discovered in only one or few samples) and common deletions (that were detected in many samples); this data reflects similar results that have been observed in the MitoBreak database where some patients' deletions are novel and have not been previously described while other patients' deletions have been identified previously in several studies (14). Two of our 'high impact' deletions were previously described in the MitoBreak database; these were the two most common deletions identified on this list, which further supports our decision to rank and prioritize deletions based on sample frequency.

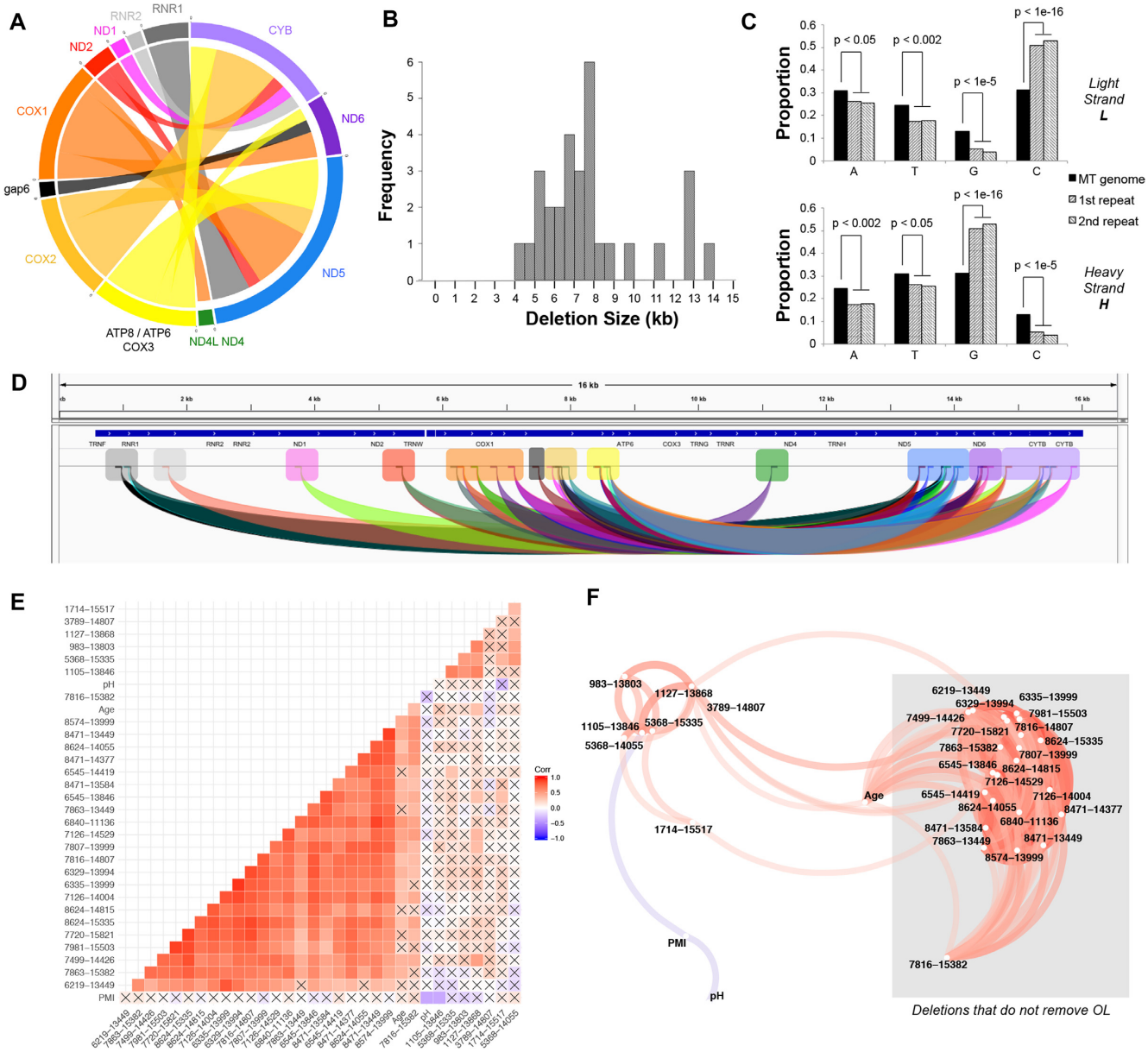

Figure 4. Characterization of the 30 most frequent mtDNA deletions. (**A**) Chord diagrams display gene involvement for the 30 most frequent mtDNA deletions. Gene regions shown are not to scale with mitochondrial coordinates; the size of the gene shown represents the total number of deletion events (5′ breakpoints + 3′ breakpoints) within that gene. Ribbons are colored based on which gene contained the 5′ breakpoint. (**B**) Histogram of deletion sizes for the 30 most frequent mtDNA deletions. (**C**) The proportion of each based used in the imperfect repeat sequences associated with the 30 deletion breakpoints compared to the entire mitochondrial genome, and results of *Z* score tests on proportions. (**D**) Condensed view of the artificial junctions.bed file generated for the 30 most frequent deletions (Supplemental Data S13) using the Integrative Genomics Viewer (IGV). (**E**) Pearson's correlation and (**F**) network analysis of significant correlations between the 30 most frequent deletions, and pH, PMI and donor age. OL (lagging strand origin-of-replication).

Tissue, age and paired analyses

We identified a significant difference in cumulative deletion read % for the different tissue types evaluated after correcting for benchmark coverage and age (Figure 5A). Peripheral samples of whole blood (BLOOD; $n = 9$) had the lowest cumulative deletion read % of 0.84 ± 0.84% (mean ± SD; range 0.15–2.56%), followed by the buccal epithelial cells (SALIVA; $n = 5$) at 2.79 ± 1.74% (mean ± SD; range 1.59–5.84%). All the brain samples examined had higher cumulative deletion read %'s than these peripheral samples: the putamen (PUT; $n = 3$) at 12.60 ± 2.47% (mean ± SD; range 11.11–15.45%), the dorsolateral prefrontal cortex (DLPFC; $n = 41$) at 21.01 ± 16.55% (mean ± SD; range 2.59–93.10%), the anterior cingulate cortex (ACC; $n = 35$) at 22.57 ± 15.68% (mean ± SD; range 4.54–90.94%), the hippocampus (HIPP; $n = 3$) at 34.22 ± 2.43% (mean ± SD; range 31.56–36.33%), and finally the caudate nucleus (CAUN; $n = 2$) at 61.21 ± 56.92% (mean ± SD; range 20.96–101.47%). The CAUN, however, only included two samples, one of which was a significant outlier with a very

Table 2. The 13 'high impact' mtDNA deletions. The position (5′-3′ breakpoints and gene), repeat sequences, sample frequencies, Sanger sequencing validation results, deletion read % and disease the deletion was detected in are shown (14). Deletions are sorted based deletion read %. All 'high impact' deletions (i.e., those with a deletion read rate ≥ 5%) detected in this study are shown. ACC (anterior cingulate cortex); DLPFC (dorsolateral prefrontal cortex); CAUN (caudate nucleus); ADO (alcohol/drug abuse/other psychiatric symptoms); BD (bipolar disorder); SZ (schizophrenia/schizoaffective disorder); MDD (major depressive disorder). Sanger sequence flanking each breakpoint and primers used for deletion validation are shown in Supplemental Figure S5

No.	Mitochondrial Deletion	Repeats and Breakpoints	No. Samples			Sanger Validation		MitoBreak Database		Relative Read Rate
	MapSplice Breakpoint 5' - 3'	imperfect repeats shown perfect repeat MapSplice break	Total n=93	Brain n=84	Blood n=9	PCR prod yes / no	gDNA yes / no	Database yes / no	Position (adj.) 5'- 3'	Deletion Read %, Diagnosis, Brain Region
1	1243 - 15340 (RNR1 - CYB)	CCACCTCCTT CCACCTCCTA	6	6	0	yes	yes	no	1243 -15340	90.1% and 85.0% for 1x MDD subject's DLPFC and ACC, respectively
2	7863 - 15617 (COX2 - CYB)	CCTCCCTTACC CGTCC TTGCC	3	3	0	yes	yes	no	7867 - 15621	52.4% for 1x MDD subject's CAUN
3	2973 - 15573 (RNR2 - CYB)	CAACAAT *no repeat* *no repeat* TCGCCTA	1	1	0	yes	yes	no	2973 - 15573	31.8% for 1x SZ subject's DLPFC
4	7816 - 14807 (COX2 - CYB)	TCATCGCCCTCCCATCCC TCATCGACCTCCCCACCC	86	84	2	yes	yes	yes	7814 - 14805	26.5% for 1x MDD subject's CAUN
5	5897 - 15840 (TRNY/COX1* - CYB)	CCTCA CCCC CCT AA T CC T	1	1	0	yes	yes	no	5897 - 15840	23.0% for 1x BD subject's DLPFC
6	1148 - 15607 (RNR1 - CYB)	AACA CTACGAG AACAAACTAGGAG	1	1	0	yes	yes	no	1148 - 15607	16.8% for 1x SZ subject's DLPFC
7	870 - 14774 (RNR1 - CYB)	ACTAACCCCA ACTAACCCCC	8	8	0	yes	yes	no	870 - 14774	10.2% for 1x MDD subject's CAUN
8	6468 - 15600 (COX1 - CYB)	GATCCGTCC TAA GATCCGTCCCTAA	50	50	0	yes	yes	yes	6468 - 15600	8.5% for 1x MDD subject's CAUN
9	571 - 15662 (D-LOOP - CYB)	CCCCCCA CCCCATC	1	1	0	yes	yes	no	571 - 15662	7.2% for 1x ADO subject's DLPFC
10	1922 - 15782 (RNR2 - CYB)	AACCAGACGAGCTACC AACCAGTA AGCTACC	45	45	0	yes	yes	no	1923 - 15783	6.8% for 1x ADO subject's ACC
11	560 - 15530 (D-LOOP - CYB)	CCAAACCCCA CC AACCCCT	13	13	0	yes	no	no	560 - 15530	6.0% for 1x ADO subject's ACC
12	488 - 14329 (D-LOOP - ND6)	ATACAACCCCCGCCC ATCCTAC CCACAACCACCACCCC ATCATAC	12	12	0	yes	no	no	488 - 14329	5.0% for 1x MDD subject's DLPFC
13	467 - 14122 (D-LOOP - ND5)	CCTCCCACTCC CTTCCCACTCA	44	44	0	yes	no	no	467 - 14122	5.0% for 1x BD subject's DLPFC

high cumulative deletion read % (Figure 5A). We also identified a significant difference in the number of unique deletions detected (per 10K coverage) after correcting for coverage and age (Figure 5B). Again, BLOOD had the lowest number of deletions with 36.4± 20.1 (mean ± SD; range 15.3–78.4 deletions), followed by the CAUN with 118.1 ± 60.7 (mean ± SD; range 75.1–161.0 deletions), the HIPP with 118.5 ± 36.9 (mean ± SD; range 80.0–153.6 deletions), the SALIVA with 118.5 ± 45.6 (mean ± SD; range 47.7–171.9 deletions), the PUT with 170.0 ± 19.7 (mean ± SD; range 153.2–191.8 deletions), the DLPFC with 195.1 ± 100.3 (mean ± SD; range 30.3–481.2 deletions), and finally the ACC with 205.1 ± 99.3 (mean ± SD; range 47.0–508.6 deletions). Lastly, we also detected a significant difference in the % burden per deletion after correcting for coverage and age (Figure 5C).

We discovered a significant and positive correlation between brain regions (ACC versus DLPFC; $n = 35$) for all three of the cumulative deletion metrics, but no significant correlation between brain and whole blood (DLPFC versus BLOOD; $n = 9$) (Figure 5D–F). Analysis of the cumulative deletion read % in paired brain regions from the same subjects demonstrated a significant and positive correlation between these cortical regions (Pearson's r: 0.842; $P = 2.2e^{-10}$), but no correlation in cumulative deletion read % in paired blood and brain samples (Pearson's r: –0.495; $P = 0.175$) (Figure 5D). Likewise, analysis of the number of deletions detected (per 10k coverage) demonstrated a significant and positive correlation between these cortical regions (Pearson's r: 0.769; $P = 6.8e^{-8}$), but no correlation in paired blood and brain samples (Pearson's r: –0.179; $P = 0.644$) (Figure 5E). Lastly, analysis of the % burden per deletion demonstrated a significant and positive correlation between these cortical regions (Pearson's r: 0.871; $P = 9.9e^{-12}$), but no correlation in cumulative deletion read % in paired blood and brain samples (Pearson's r: –0.198; $P = 0.609$) (Figure 5F). The correlation tests for both the cumulative deletion % and the % burden per deletion between brain regions were influenced by a single outlier subject that had a 'high impact' deletion in both the ACC and DLPFC; however, these cortical brain regions were still significantly correlated if this subject was removed (cumulative deletion read %; Pearson's r: 0.645; $P = 3.7e^{-5}$) (% burden per deletion; Pearson's r: 0.460; $P = 0.006$).

Our final set of analyses revealed several significant and positive correlations between the brain regions' cumulative deletion metrics and subject age, but no significant correlations between age and the deletion levels detected in blood (Figure 5G–I). Linear regression models between the cumulative deletion metrics and age were evaluated after correcting for coverage ± pH and RIN. Analysis of the cumulative deletion read % demonstrated a significant, pos-

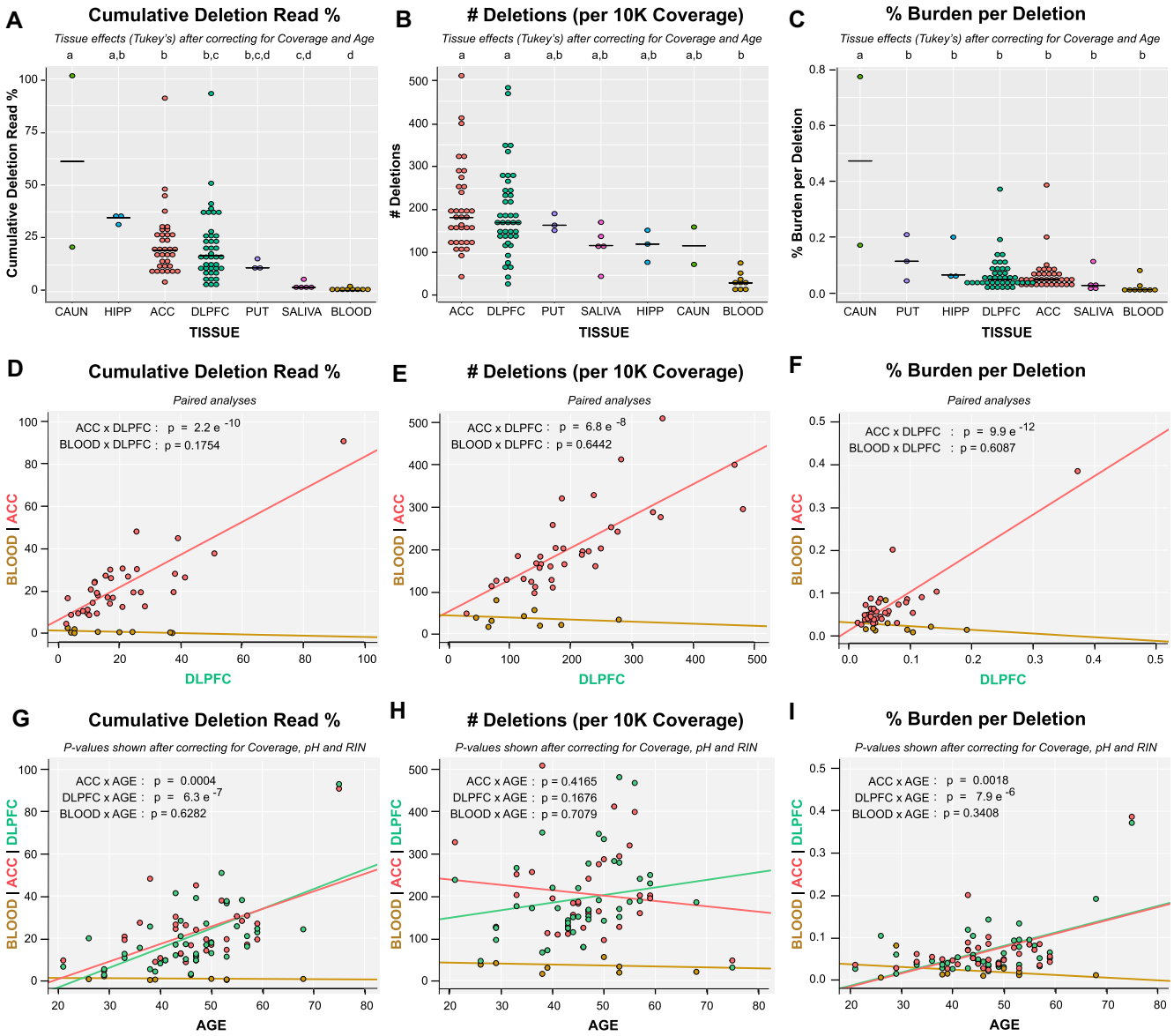

Figure 5. Tissue, age and paired analyses. Tissue analyses of (**A**) the cumulative deletion read %, (**B**) number of deletion species detected (per 10k coverage) and (**C**) the % burden per deletion. Statistical results shown (letters above tissue groups) are from Tukey's post-hoc analyses after correcting for coverage and age. Paired analyses of subject-matched brain regions (ACC x DLPFC, $n = 35$) and tissues (DLPFC × BLOOD, $n = 9$) for (**D**) the cumulative deletion read %, (**E**) number of deletion species detected (per 10k coverage) and (**F**) the % burden per deletion. Statistical results shown are from Pearson's correlation tests. Age analysis of (**G**) the cumulative deletion read %, (**H**) number of deletion species detected (per 10k coverage) and (**I**) the % burden per deletion in ACC (anterior cingulate cortex, $n = 35$), DLPFC (dorsolateral prefrontal cortex, $n = 41$), and BLOOD (whole blood, $n = 9$). Statistical results shown are linear regression models after correcting for coverage, pH and RIN.

itive correlation with age for both the DLPFC ($n = 41$; Adj. R^2 of model: 0.591; Age P-value: $6.3e^{-7}$) and ACC ($n = 35$; Adj. R^2 of model: 0.400; Age P-value: 0.004), but not for BLOOD ($n = 9$; Adj. R^2 of model: –0.012; Age P-value: 0.628). In contrast, analysis of the number of deletions (per 10k coverage) did not show a significant correlation with age in any of the tissues evaluated (ACC, DLPFC and BLOOD) (Figure 5H). Lastly, analysis of the % burden per deletion demonstrated a significant and positive correlation with age for both the DLPFC ($n = 41$; Adj. R^2 of model: 0.550; Age P-value: $7.9e^{-6}$) and ACC ($n = 35$; Adj. R^2 of model: 0.312; Age P-value: 0.0018), but not for BLOOD ($n = 9$; Adj. R^2 of model: 0.115; Age P-value: 0.341) (Figure 5I).

In order to assess if these biological findings would be lost or retained using a different mtDNA deletion detection approach, we have reproduced Figure 5 in its entirety but with results for the 93 samples obtained from MitoDel (using MapSplice alignment for alignment consistency) instead of Splice-Break (Supplemental Figure S14). Results for the cumulative deletion read % were largely identical between methods (Figure 5A, D and G versus Supplemental Figure S14A, D and G). This result was also reflected in our highly significant Pearson's correlation analysis of the cumulative

deletion read % between Splice-Break and MitoDel (Pearson's *r*: 0.950; $P = 2.98e^{-50}$) (Supplemental Figure S14J). Results for the number of unique deletions detected (per 10k coverage) after correcting for coverage and age were quite different between methods, however, with ∼10-fold more deletion species being detected with Splice-Break versus MitoDel (Figure 5B, E and H versus Supplemental Figure S14B, E and H). This had an effect on the ordering of which brain regions had the most/least deletion species, and analysis of paired brain regions (ACC × DLPFC) was much less significant using MitoDel (Supplemental Figure S14E). Again, Pearson's correlation of the number of unique deletions detected (per 10k coverage) between Splice-Break and MitoDel was significant (Pearson's *r*: 0.392; $P = 6.64e^{-5}$), albeit to a lesser extent than the cumulative deletion read % (Supplemental Figure S14J and K).

Disease effects

We evaluated the cumulative deletion read % (Figure 6) and number of deletions (per 10k coverage) for disease effects in the ACC and DLPFC after correcting for coverage and subject age. Subjects were grouped based on primary diagnosis, and disease effects were tested across 9 subjects with Major Depressive Disorder (MDD), 10 subjects with Bipolar Disorder (BD), 12 subjects with Schizophrenia or Schizoaffective Disorder (SZ), 8 subjects with none of the above diagnoses but a history of alcohol dependence, drug abuse, and possibly other mental health issues (ADO), and 2 healthy controls (CTRL) with no history of psychiatric symptoms or alcohol/drug abuse (Supplemental Table S1). We identified no significant difference in the cumulative mtDNA deletion read % in the ACC ($P = 0.518$) or in the DLPFC ($P = 0.338$) across disorders (Figure 6A and B). Likewise, our initial analysis of the DLPFC/ACC ratio for the cumulative deletion read % did not reveal a significant difference across disorders ($P = 0.128$) (Figure 6C). However, this data did appear intriguing in that the SZ and BD subjects had a higher DLPFC/ACC ratio than the CTRL, ADO or MDD subjects, so we performed a post-hoc analysis on these pooled groups (Figure 6D). When combined, the BD + SZ group had a significantly higher DLPFC/ACC cumulative deletion read % ratio ($P = 0.009$) than the MDD + ADO + CTRL group (Figure 6D). We saw no significant disease effects in the number of unique deletions (per 10k benchmark coverage) identified in the ACC ($P = 0.065$) or DLPFC ($P = 0.566$) nor in the DLPFC/ACC ratio ($P = 0.879$) (data not shown).

While the above analyses were important to look for group mean differences across psychiatric disorders, we believe that performing an 'analysis of outliers' is equally as important when investigating a heterogeneous cohort where a small subset of subjects may be predicted to have a mitochondrial pathology. In addition, we compared our data to subjects with known mitochondrial pathologies, specifically Kearns–Sayre syndrome (KSS) muscle and Pearson syndrome (PS) blood (Figure 6E–K), in order to assess if any outlier we identified had mtDNA deletions at or above the levels observed in 'hallmark disorders'. Using a cut-off of 1.5-times the interquartile range (IQR) of all brain samples, an outlier would be defined as a sample with a cumulative deletion read % of 52.01 or more; however, we used a more conservative threshold of 78.1% for this analysis of outliers, which we defined as 2-fold more than the 95th percentile of all brain samples not be considered outliers by the IQR rules. We performed this analysis across all available brain regions, and identified two MDD subjects (three brain tissue samples) that met our threshold to be called an outlier (Figure 6I–K). These samples also displayed smaller LR PCR band(s) on the agarose gel (Supplemental Figure S2), and exhibited a sharp 'shelf-like' drop in coverage at the mtDNA deletion breakpoints (Figure 6I, J and Supplemental Figure S6). One MDD subject was from our SBB cohort; this subject had the No. 1 'high impact' deletion (1243–15340) at deletion read %'s of 90.1% (DLPFC) and 85.0% (ACC) (Table 2). This single deletion made up the vast majority of the cumulative deletion read %'s identified in this subject, which were 93.1 and 90.9% for the DLPFC and ACC, respectively (Figure 6J–K). This subject was a 75-year-old male with MDD and diabetes mellitus, which we believe is a clinical anecdote worth mentioning given its association with mitochondrial pathologies and deletions, specifically (5,8,9,11,42–45). The other MDD subject was from our UCI Brain Bank cohort; this subject was a 46-year-old male with MDD who committed suicide. Five brain regions (ACC, DLPFC, HIPP, PUT, CAUN) and whole blood (BLOOD) were evaluated in this subject, but only the caudate nucleus (CAUN) of this subject was identified as an outlier (Figure 6I and K); this MDD subject's CAUN had the No. 2, 4, 7 and 8 'high impact' deletions at deletion read %'s of 52.4%, 26.5%, 10.2% and 8.5% (Table 2). Cumulatively, these four clonally expanded deletions had a deletion read % of 97.6%, which (like the other subject) made up the vast majority of the cumulative deletion read % (101.5%) identified in this sample (Figure 6i and k). This result is particularly intriguing because it provides evidence that clonally expanded mtDNA deletions can be found exclusively in specific brain regions in humans while being absent (or at low levels) in other cortical regions.

We performed a hypergeometric mean distribution test on the proportion of MDD subjects that had a brain region identified as an outlier (2/9) in comparison to the entire cohort of subjects (2/41), and identified a significant ($P = 0.044$) enrichment in MDD (Figure 6K). This data supports a link between mtDNA deletions and depression, specifically suggesting a subset or subgroup of MDD cases may be due to an mtDNA deletion pathology in the brain (46). Based on the mtDNA deletion profiles we observed in these two distinctive MDD subjects, we suspect the 'high impact' deletion(s) occurred and/or were clonally expanded early during brain development, although we speculate they may not have accrued a high enough rate to supersede a pathological threshold until later in life. This disease analysis is particularly intriguing as there is evidence from both mouse and human studies suggesting depression is a phenotype associated with mitochondrial dysfunction. Specifically, transgenic mice that express mutant *POLG* and accumulate mtDNA deletions in the brain have been characterized as having 'mood disorder-like' phenotypes and 'depression-like episodes' that include distorted activity patterns in diurnal rhythmicity and may reflect the circadian rhythm dysfunctions observed in MDD (47–49).

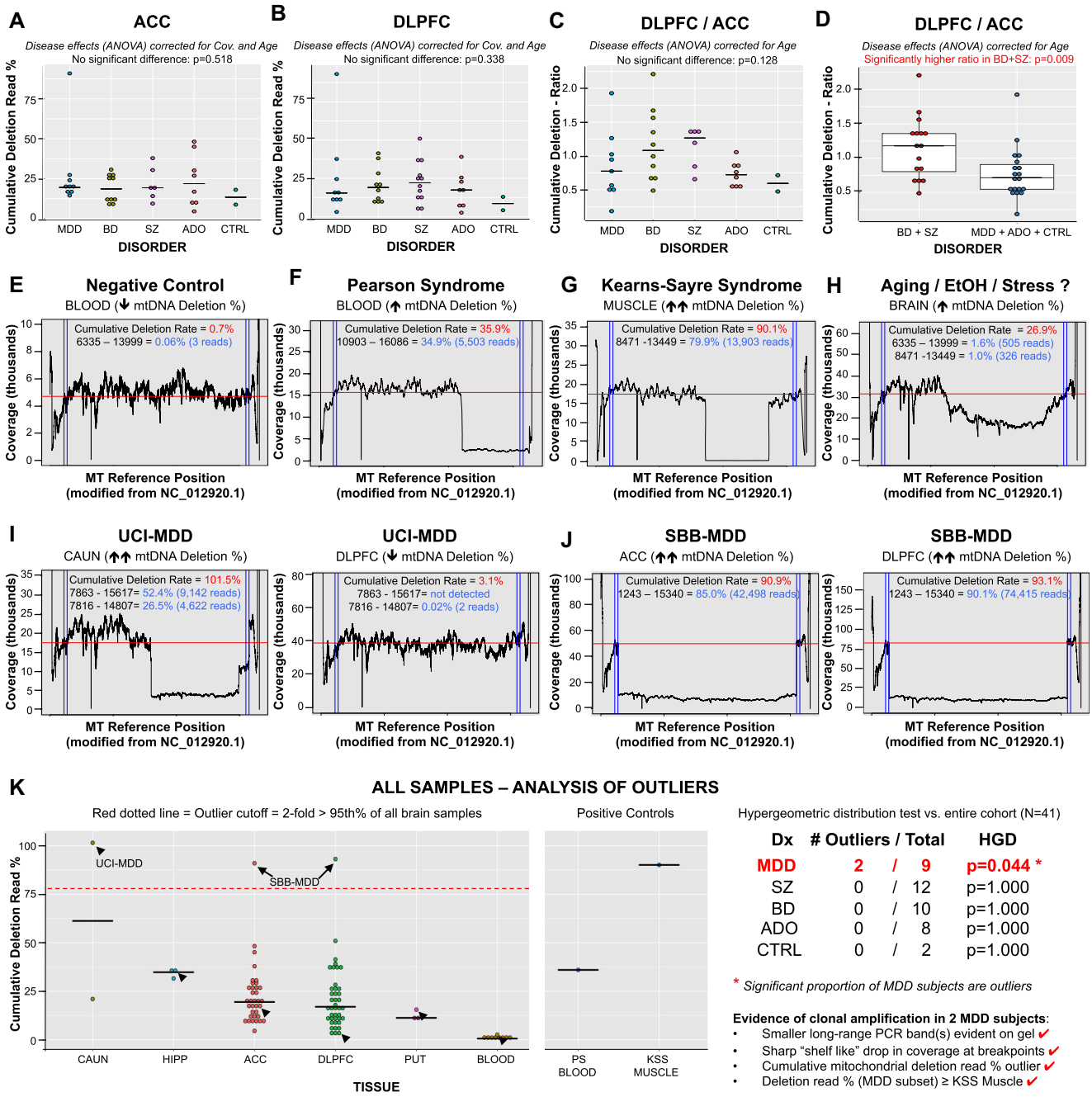

Figure 6. Psychiatric disorder effects and comparisons to mitochondrial pathologies. Cumulative deletion read % across diagnoses in the (**A**) ACC (anterior cingulate cortex) and (**B**) DLPFC (dorsolateral prefrontal cortex), (**C**) the ratio of these two brain regions, and (**D**) pooled analysis of (C) after combining BD and SZ groups. Splice-Break coverage plots of mtDNA from (**E**) a typical blood sample with very few deletions, (**F**) a blood sample from a subject with Pearson's Syndrome, (**G**) a muscle sample from a subject with Kearns–Sayre syndrome, (**H**) an aging/adult brain sample with many low level mtDNA deletions (but no single, clonally amplified deletion), (**I**) two brain regions from one MDD subject, where the caudate nucleus (CAUN) displayed a large, clonally amplified deletion but the dorsolateral prefrontal cortex (DLPFC) did not and (**J**) two brain regions from another MDD subject, where both the anterior cingulate cortex (ACC) and DLPFC were affected by the same clonally amplified deletion. (**K**) Analysis of outliers based on threshold criteria (>2-fold more than 95th percentile of all other brain regions). Arrowheads point to one MDD subject where only the caudate nucleus (CAUN) was identified as an outlier, while all other brain regions from this subject were considered normal; full arrows denote one MDD subject where both brain regions analyzed (ACC and DLPFC) were identified as outliers. Positive controls of PS blood and KSS muscle are shown alongside for comparison. Statistics shown for are from one-way ANOVA (A–G) or hypergeometric mean distribution tests.

Likewise, both children and adult patients with diagnosed mitochondrial diseases have displayed concurrent psychiatric symptoms that often include depression (46,47,50–54). The two MDD subjects (three brain tissue samples) we identified as outliers also had cumulative deletion read %'s greater than KSS muscle (90.1%; Figure 6G and K) or PS blood (35.9%; Figure 6F and K), which further supports our hypothesis that these mtDNA deletions are at levels high enough to cause mitochondrial dysfunction, cellular pathology and ultimately brain circuit dysfunction (i.e., depression).

CONCLUSIONS

This study demonstrates the efficacy and reliability of the Splice-Break pipeline to detect and quantify mtDNA deletions. Splice-Break can be used to study large datasets of individual mtDNA deletions rates as well as cumulative deletion metrics with respect to tissue, brain region, age and disease. The catalogue of 4489 mtDNA deletions we describe here only includes those with breakpoints between positions 357–15925 (NC_012920.1) of the mitochondrial genome as we filtered out the majority of the control region in this analysis. Future studies may include analysis of mtDNA deletion breakpoints in the D-loop and extended control region, which includes a 3′ breakpoint 'hotspot' at position 16071, perhaps using different primers (with different binding positions) for the LR PCR (7,14,32). Indeed any analysis of mtDNA deletions that utilizes LR PCR will be limited to only discover molecules that can be successfully amplified with the primers chosen. Interestingly, the 'common deletion' (MapSplice 8471–13449; adjusted 8482–13460) was actually the third most common deletion in this collective dataset and was detected in 85/93 (91.4%) samples at a deletion read % of 0.37 ± 0.32% (mean ± SD). The most frequent deletion we detected (MapSplice 6335–13999; adjusted 6341–14005) was observed in 92/93 (98.9%) biological samples at a deletion read % of 0.65 ± 0.57% (mean ± SD). This latter deletion has additionally been confirmed as the most frequent mtDNA deletion detected by the Splice-Break pipeline using an independent group of ∼90 brain samples (data not shown).

We observed a significant correlation ($P = 3.43e^{-17}$) between the qPCR results of the 4977 bp 'common deletion' and our Splice-Break results for this mtDNA deletion. Using artificial data of five different mtDNA deletion sequences, we observed that the test sensitivity of our pipeline can vary between deletions (range 38.8–74.6%), which we believe is a reflection on how MapSplice handles reads that contain the various repeat sequences associated with deletion breakpoints. We also observed a high level of test specificity for these five mtDNA deletions (range 93.2–100%), and were able to optimize our filtering parameters so that low abundance deletions (i.e. 1% heteroplasmy) in particular would be called with 100% specificity. Analysis of a complex artificial file that had 60 deletions combined together also had a high detection sensitivity (71.2%) that we were unable to match with the other methods tested. Taken together, these results demonstrate that our Splice-Break pipeline can quantify the *relative* levels of mtDNA deletions with high confidence, but our deletion read % described should not be interpreted as a direct measure of heteroplasmy rate due to factors of PCR amplification bias and deletion-specific differences in test sensitivity. We also observed a significant correlation ($P = 2.98e^{-50}$) in the cumulative deletion read % determined by our Splice-Break pipeline and the MitoDel analysis tool (when the same alignment algorithm was used for both), but detected ∼10-times more deletion species with Splice-Break and a more significant correlation between paired brain regions. These methods comparisons are encouraging and suggest our biological conclusions would largely remain the same with the usage of either analysis tool (if an RNA-Seq aligner was used for read mapping); however, our Splice-Break pipeline is particularly well-suited for the analysis of homogenate tissue where many deletion species are predicted to occur concurrently at a low rate.

We observed significantly more deletions in brain than in blood using all three of our cumulative deletion metrics- these results were expected and the analysis was performed largely for assessment of the Spice-Break pipeline. This discrepancy in tissue differences is also not surprising given that adolescent and adult patients with diagnosed mitochondrial deletion pathologies often do not have observable mtDNA deletion loads in their blood even though their muscles are highly affected (1,2,11). Our results further suggest that mtDNA deletion levels observed in peripheral whole blood are not good biomarkers for predicting what the mtDNA deletion levels are in the brain; however, there are significant correlations between brain regions that may have some value for forensic studies. We also observed a significant and positive correlation between subject age and the cumulative deletion read % observed in the brain (ACC or DLPFC), but did not detect a significant correlation between subject age and the deletion levels observed in blood. This corroborates previous reports that have observed a progressive accumulation of mtDNA deletions and mutations in post-mitotic (non-dividing) tissues, but a loss (or purging) of these aberrant mitochondria in blood with advanced age due to the rapid division of leukocytes (11,12). The age correlations we observed in the brain will be interesting to investigate with regards to late-onset neurological disorders such as Alzheimer's disease, Parkinson's disease and mild cognitive impairment, in addition to complex disorders with symptoms outside of the central nervous system such as chronic fatigue syndrome, diabetes, and age-related eye diseases.

Finally, it will be worthwhile to pursue these mtDNA findings with Splice-Break in greater depth in multiple brain regions using a larger cohort of individuals with psychiatric disorders and controls. We found that clonally expanded, 'high impact' deletions can be present at high levels in one brain region, but remain absent or at very low levels in other regions from the same subject. This result indicates that screening efforts for mtDNA deletion loads in the brain may require multiple brain regions in order to see the full picture. In addition, analysis of mtDNA deletions in specific cell populations (e.g. neurons versus glia, in specific neuron types and/or cortical layers) will be important follow on studies; however, the fact that we were able to detect significant differences and correlations with frozen, homogenate brain tissue is encouraging and may provide a more high-

throughput and cost-effective approach for an initial assessment of disease and aging. The highest and most overt deletion burdens we observed occurred in two MDD subjects, one of which had only the caudate nucleus affected with 'high impact' mtDNA deletions. This region, in particular, should be examined in a larger cohort of subjects with psychiatric diagnoses, and should be evaluated for possible correlations with depression rating scales and other prognostic indicators. We suggest that future mtDNA deletion screening efforts focused on depression should evaluate multiple brain regions within the cortico–striatal–thalamic loop circuits. These circuits include all of the brain regions that we observed significant mtDNA deletion accumulation in, and are important for cognition, emotion, and motor control (55,56).

DATA AVAILABILITY

The catalogue of 4489 mtDNA deletions can be found in Supplemental Data S12, and updated versions of this list will be provided on the MitoBreak website: http://mitobreak.portugene.com. Raw sequence data (unaligned, paired-end FASTQ files) for all 93 postmortem samples described in this study can be downloaded from GEO (GSE118615).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are greatly appreciative to all of the autopsy donors and their families, without whom this work would not be possible.

FUNDING

Della Martin Foundation [to W.B., B.H. postdoctoral fellowship]; National Institute of Mental Health (NIMH), National Institutes of Health [R01MH08580, R21MH099440 to M.V. and R01MH097082 to A.S.]; Portuguese Foundation for Science and Technology [IF/01356/2012 to F.P.]. Funding support for postmortem brains collected at the Southwest Brain Bank at Texas Tech University Health Sciences Center (TTUHSC) was provided by the McKee Foundation and TTUHSC-El Paso. Funding support for postmortem brains collected at the University of California-Irvine (UCI) Brain Bank was provided by the Pritzker Neuropsychiatric Disorders Research Consortium. Funding for open access charge: National Institute of Mental Health.
Conflict of interest statement. None declared.

REFERENCES

1. Holt,I.J., Harding,A.E. and Morgan-Hughes,J.A. (1988) Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature*, **331**, 717–719.
2. Lestienne,P. and Ponsot,G. (1988) Kearns–Sayre syndrome with muscle mitochondrial DNA deletion. *Lancet*, **1**, 885.
3. Zeviani,M., Moraes,C.T., DiMauro,S., Nakase,H., Bonilla,E., Schon,E.A. and Rowland,L.P. (1988) Deletions of mitochondrial DNA in Kearns–Sayre syndrome. *Neurology*, **38**, 1339–1346.
4. Rotig,A., Cormier,V., Koll,F., Mize,C.E., Saudubray,J.M., Veerman,A., Pearson,H.A. and Munnich,A. (1991) Site-specific deletions of the mitochondrial genome in the Pearson marrow-pancreas syndrome. *Genomics*, **10**, 502–504.
5. Ballinger,S.W., Shoffner,J.M., Hedaya,E.V., Trounce,I., Polak,M.A., Koontz,D.A. and Wallace,D.C. (1992) Maternally transmitted diabetes and deafness associated with a 10.4 kb mitochondrial DNA deletion. *Nat. Genet.*, **1**, 11–15.
6. Yamashita,S., Nishino,I., Nonaka,I. and Goto,Y. (2008) Genotype and phenotype analyses in 136 patients with single large-scale mitochondrial DNA deletions. *J. Hum. Genet.*, **53**, 598–606.
7. Damas,J., Samuels,D.C., Carneiro,J., Amorim,A. and Pereira,F. (2014) Mitochondrial DNA rearrangements in health and disease—a comprehensive study. *Hum. Mutat.*, **35**, 1–14.
8. DiMauro,S. and Hirano,M. (1993) Mitochondrial DNA Deletion Syndromes. In: Adam,MP, Ardinger,HH, Pagon,RA, Wallace,SE, Bean,LJH, Mefford,HC, Stephens,K, Amemiya,A and Ledbetter,N (eds). *GeneReviews((R))*. Seattle.
9. Taylor,R.W. and Turnbull,D.M. (2005) Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.*, **6**, 389–402.
10. Tuppen,H.A., Blakely,E.L., Turnbull,D.M. and Taylor,R.W. (2010) Mitochondrial DNA mutations and human disease. *Biochim. Biophys. Acta*, **1797**, 113–128.
11. Pitceathly,R.D., Rahman,S. and Hanna,M.G. (2012) Single deletions in mitochondrial DNA–molecular mechanisms and disease phenotypes in clinical practice. *Neuromuscul. Disord.*, **22**, 577–586.
12. Stewart,J.B. and Chinnery,P.F. (2015) The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.*, **16**, 530–542.
13. Grady,J.P., Campbell,G., Ratnaike,T., Blakely,E.L., Falkous,G., Nesbitt,V., Schaefer,A.M., McNally,R.J., Gorman,G.S., Taylor,R.W. *et al.* (2014) Disease progression in patients with single, large-scale mitochondrial DNA deletions. *Brain*, **137**, 323–334.
14. Damas,J., Carneiro,J., Amorim,A. and Pereira,F. (2014) MitoBreak: the mitochondrial DNA breakpoints database. *Nucleic Acids Res.*, **42**, D1261–D1268.
15. El-Hattab,A.W., Craigen,W.J. and Scaglia,F. (2017) Mitochondrial DNA maintenance defects. *Biochim. Biophys. Acta*, **1863**, 1539–1555.
16. Milone,M. (2013) Mitochondrial DNA multiple deletion syndromes, autosomal dominant and recessive (POLG, POLG2, TWINKLE, and ANT1). In: Wong,LC (ed). *Mitochondrial Disorders Caused by Nuclear Genes*. Springer, NY, pp. 123–140.
17. Spelbrink,J.N., Li,F.Y., Tiranti,V., Nikali,K., Yuan,Q.P., Tariq,M., Wanrooij,S., Garrido,N., Comi,G., Morandi,L. *et al.* (2001) Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nat. Genet.*, **28**, 223–231.
18. Paramasivam,A., Meena,A.K., Pedaparthi,L., Jyothi,V., Uppin,M.S., Jabeen,S.A., Sundaram,C. and Thangaraj,K. (2016) Novel mutation in C10orf2 associated with multiple mtDNA deletions, chronic progressive external ophthalmoplegia and premature aging. *Mitochondrion*, **26**, 81–85.
19. Leshinsky-Silver,E., Malinger,G., Ben-Sira,L., Kidron,D., Cohen,S., Inbar,S., Bezaleli,T., Levine,A., Vinkler,C., Lev,D. *et al.* (2011) A large homozygous deletion in the SAMHD1 gene causes atypical Aicardi-Goutieres syndrome associated with mtDNA deletions. *Eur. J. Hum. Genet.*, **19**, 287–292.
20. Pfeffer,G., Gorman,G.S., Griffin,H., Kurzawa-Akanbi,M., Blakely,E.L., Wilson,I., Sitarz,K., Moore,D., Murphy,J.L., Alston,C.L. *et al.* (2014) Mutations in the SPG7 gene cause chronic progressive external ophthalmoplegia through disordered mitochondrial DNA maintenance. *Brain*, **137**, 1323–1336.
21. Gorman,G.S., Pfeffer,G., Griffin,H., Blakely,E.L., Kurzawa-Akanbi,M., Gabriel,J., Sitarz,K., Roberts,M., Schoser,B., Pyle,A. *et al.* (2015) Clonal expansion of secondary mitochondrial DNA deletions associated with spinocerebellar ataxia type 28. *JAMA Neurol.*, **72**, 106–111.
22. Cree,L.M., Samuels,D.C. and Chinnery,P.F. (2009) The inheritance of pathogenic mitochondrial DNA mutations. *Biochim. Biophys. Acta*, **1792**, 1097–1102.

23. Moraes,C.T., Atencio,D.P., Oca-Cossio,J. and Diaz,F. (2003) Techniques and pitfalls in the detection of pathogenic mitochondrial DNA mutations. *J. Mol. Diagn.*, **5**, 197–208.
24. Krishnan,K.J., Bender,A., Taylor,R.W. and Turnbull,D.M. (2007) A multiplex real-time PCR method to detect and quantify mitochondrial DNA deletions in individual cells. *Anal. Biochem.*, **370**, 127–129.
25. Belmonte,F.R., Martin,J.L., Frescura,K., Damas,J., Pereira,F., Tarnopolsky,M.A. and Kaufman,B.A. (2016) Digital PCR methods improve detection sensitivity and measurement precision of low abundance mtDNA deletions. *Sci. Rep.*, **6**, 25186.
26. He,L., Chinnery,P.F., Durham,S.E., Blakely,E.L., Wardell,T.M., Borthwick,G.M., Taylor,R.W. and Turnbull,D.M. (2002) Detection and quantification of mitochondrial DNA deletions in individual cells by real-time PCR. *Nucleic Acids Res.*, **30**, e68.
27. Grady,J.P., Murphy,J.L., Blakely,E.L., Haller,R.G., Taylor,R.W., Turnbull,D.M. and Tuppen,H.A. (2014) Accurate measurement of mitochondrial DNA deletion level and copy number differences in human skeletal muscle. *PLoS One*, **9**, e114462.
28. Harbottle,A., Krishnan,K.J. and Birch-Machin,M.A. (2004) Implications of using the ND1 gene as a control region for real-time PCR analysis of mitochondrial DNA deletions in human skin. *J. Invest. Dermatol.*, **122**, 1518–1521.
29. Guo,Y., Li,J., Li,C.I., Shyr,Y. and Samuels,D.C. (2013) MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics*, **29**, 1210–1211.
30. Seneca,S., Vancampenhout,K., Van Coster,R., Smet,J., Lissens,W., Vanlander,A., De Paepe,B., Jonckheere,A., Stouffs,K. and De Meirleir,L. (2015) Analysis of the whole mitochondrial genome: translation of the Ion Torrent Personal Genome Machine system to the diagnostic bench? *Eur. J. Hum. Genet.*, **23**, 41–48.
31. Bosworth,C.M., Grandhi,S., Gould,M.P. and LaFramboise,T. (2017) Detection and quantification of mitochondrial DNA deletions from next-generation sequence data. *BMC Bioinformatics*, **18**, 407.
32. Zambelli,F., Vancampenhout,K., Daneels,D., Brown,D., Mertens,J., Van Dooren,S., Caljon,B., Gianaroli,L., Sermon,K., Voet,T. *et al.* (2017) Accurate and comprehensive analysis of single nucleotide variants and large deletions of the human mitochondrial genome in DNA and single cells. *Eur. J. Hum. Genet.*, **25**, 1229–1236.
33. Goudenège,D., Bris,C., Hoffmann,V., Desquiret-Dumas,V., Jardel,C., Rucheton,B., Bannwarth,S., Paquis-Flucklinger,V., Lebre,A.S., Colin,E. *et al.* (2018) eKLIPse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data. *Genet. Med.*, doi:10.1038/s41436-018-0350-8.
34. Wang,K., Singh,D., Zeng,Z., Coleman,S.J., Huang,Y., Savich,G.L., He,X., Mieczkowski,P., Grimm,S.A., Perou,C.M. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.
35. Zhang,W., Cui,H. and Wong,L.J. (2012) Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. *Clin. Chem.*, **58**, 1322–1331.
36. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
37. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Sonali,J., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
38. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
39. Parada,G.E., Munita,R., Cerda,C.A. and Gysling,K. (2014) A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.*, **42**, 10564–10578.
40. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
41. Dong,D.W., Pereira,F., Barrett,S.P., Kolesar,J.E., Cao,K., Damas,J., Yatsunyk,L.A., Johnson,F.B. and Kaufman,B.A. (2014) Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. *BMC Genomics*, **15**, 677.
42. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
43. Petersen,K.F., Befroy,D., Dufour,S., Dziura,J., Ariyan,C., Rothman,D.L., DiPietro,L., Cline,G.W. and Shulman,G.I. (2003) Mitochondrial dysfunction in the elderly: possible role in insulin resistance. *Science*, **300**, 1140–1142.
44. Newsholme,P., Gaudel,C. and Krause,M. (2012) Mitochondria and diabetes. An intriguing pathogenetic role. *Adv. Exp. Med. Biol.*, **942**, 235–247.
45. Soleimanpour,S.A., Gupta,A., Bakay,M., Ferrari,A.M., Groff,D.N., Fadista,J., Spruce,L.A., Kushner,J.A., Groop,L., Seeholzer,S.H. *et al.* (2014) The diabetes susceptibility gene Clec16a regulates mitophagy. *Cell*, **157**, 1577–1590.
46. Morava,E. and Kozicz,T. (2013) Mitochondria and the economy of stress (mal)adaptation. *Neurosci. Biobehav. Rev.*, **37**, 668–680.
47. Kasahara,T., Kubota,M., Miyauchi,T., Noda,Y., Mouri,A., Nabeshima,T. and Kato,T. (2006) Mice with neuron-specific accumulation of mitochondrial DNA mutations show mood disorder-like phenotypes. *Mol. Psychiatry*, **11**, 577–593.
48. Kasahara,T., Takata,A., Kato,T.M., Kubota-Sakashita,M., Sawada,T., Kakita,A., Mizukami,H., Kaneda,D., Ozawa,K. and Kato,T. (2016) Depression-like episodes in mice harboring mtDNA deletions in paraventricular thalamus. *Mol. Psychiatry*, **21**, 39–48.
49. Li,J.Z., Bunney,B.G., Meng,F., Hagenauer,M.H., Walsh,D.M., Vawter,M.P., Evans,S.J., Choudary,P.V., Cartagena,P., Barchas,J.D. *et al.* (2013) Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 9950–9955.
50. Siciliano,G., Tessa,A., Petrini,S., Mancuso,M., Bruno,C., Grieco,G.S., Malandrini,A., DeFlorio,L., Martini,B., Federico,A. *et al.* (2003) Autosomal dominant external ophthalmoplegia and bipolar affective disorder associated with a mutation in the ANT1 gene. *Neuromuscul. Disord.*, **13**, 162–165.
51. Fattal,O., Link,J., Quinn,K., Cohen,B.H. and Franco,K. (2007) Psychiatric comorbidity in 36 adults with mitochondrial cytopathies. *CNS Spectr.*, **12**, 429–438.
52. Koene,S., Kozicz,T.L., Rodenburg,R.J., Verhaak,C.M., de Vries,M.C., Wortmann,S., van de Heuvel,L., Smeitink,J.A. and Morava,E. (2009) Major depression in adolescent children consecutively diagnosed with mitochondrial disorder. *J. Affect. Disord.*, **114**, 327–332.
53. Smits,B.W., Fermont,J., Delnooz,C.C., Kalkman,J.S., Bleijenberg,G. and van Engelen,B.G. (2011) Disease impact in chronic progressive external ophthalmoplegia: more than meets the eye. *Neuromuscul. Disord.*, **21**, 272–278.
54. Manji,H., Kato,T., Di Prospero,N.A., Ness,S., Beal,M.F., Krams,M. and Chen,G. (2012) Impaired mitochondrial function in psychiatric disorders. *Nat. Rev. Neurosci.*, **13**, 293–307.
55. Rodriguez-Oroz,M.C., Jahanshahi,M., Krack,P., Litvan,I., Macias,R., Bezard,E. and Obeso,J.A. (2009) Initial clinical manifestations of Parkinson's disease: features and pathophysiological mechanisms. *Lancet Neurol.*, **8**, 1128–1139.
56. Peters,S.K., Dunlop,K. and Downar,J. (2016) Cortico-Striatal-Thalamic loop circuits of the salience network: a central pathway in Psychiatric disease and treatment. *Front. Syst. Neurosci.*, **10**, 104.