The landscape of structural variation in aye-ayes (*Daubentonia madagascariensis*)

Cyril J. Versoza[1], Jeffrey D. Jensen[1], Susanne P. Pfeifer[1,*]

[1] Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ, USA

[*] corresponding author: susanne@spfeiferlab.org

keywords: primate; strepsirrhine; Daubentoniidae; copy number variation; structural variation; population genomics

ABSTRACT

Aye-ayes (*Daubentonia madagascariensis*) are one of the 25 most critically endangered primate species in the world. Endemic to Madagascar, their small and highly fragmented populations make them particularly vulnerable to both genetic disease and anthropogenic environmental changes. Over the past decade, conservation genomic efforts have largely focused on inferring and monitoring population structure based on single nucleotide variants to identify and protect critical areas of genetic diversity. However, the recent release of a highly contiguous genome assembly allows, for the first time, for the study of structural genomic variation (deletions, duplications, insertions, and inversions) which are likely to impact a substantial proportion of the species' genome. Based on whole-genome, short-read sequencing data from 14 individuals, >1,000 high-confidence autosomal structural variants were detected, affecting ~240 kb of the aye-aye genome. The majority of these variants (>85%) were deletions shorter than 200 bp, consistent with the notion that longer structural mutations are often associated with strongly deleterious fitness effects. For example, two deletions longer than 850 bp located within disease-linked genes were predicted to impose substantial fitness deficits owing to a resulting frameshift and gene fusion, respectively; whereas several other major effect variants outside of coding regions are likely to impact gene regulatory landscapes. Taken together, this first glimpse into the landscape of structural variation in aye-ayes will enable future opportunities to advance our understanding of the traits impacting the fitness of this endangered species, as well as allow for enhanced evolutionary comparisons across the full primate clade.

INTRODUCTION

Gaining a better understanding of the process of mutation is of fundamental importance to characterize genetic variation within and between populations and species, as well as to provide insights into both the drivers of local adaptation and the factors underlying disease. Over the past decades, the primary focus of many primate population genomic studies has been on elucidating the causes and consequences of point mutations, using single nucleotide variants (SNVs) to infer the rates and patterns of recombination, population demographic history, and natural selection (e.g., Auton et al. 2012; Simkin et al. 2014; Pfeifer and Jensen 2016; Stevison et al. 2016; Nielsen et al. 2017; Pfeifer 2017a, 2020a,b, 2021; Ghafoor et al. 2023; Johri et al. 2023; Soni et al. 2024a,b; Soni and Jensen 2024; Versoza et al. 2024a,b; Versoza, Lloret-Villas, et al. 2024). This common emphasis on SNVs, however, has resulted in a general neglect of the largest source of heritable variation, namely structural variation. Structural variants (SVs) – including copy number variants (defined here as deletions and duplications larger than 50 bp in size) and balanced rearrangements such as inversions – affect more nucleotides than SNVs in the primate genomes examined to date (Redon et al. 2006; Conrad et al. 2010; Pang et al. 2010; Sudmant et al. 2010, 2013, 2015a,b; Zarrei et al. 2015; Mao et al. 2024; and see the reviews of Conrad and Hurles 2007 and Gökçümen and Lee 2009). Moreover, due to their size, SVs often impact coding and regulatory regions which, in turn, can alter gene dosage, genome structure, or modify the timing and/or level of gene expression (Chaignat et al. 2011; Chiang et al. 2017), making SVs one of the main factors impacting phenotypic adaptation as well as disease

susceptibility (Lin and Gökçümen 2019; and see reviews by Girirajan et al. 2011, Iskow et al. 2012, and Hollox et al. 2022).

Yet, despite their importance, the landscape of structural variation remains poorly characterized in many species. This neglect largely owes to the fact that SVs are more challenging to accurately identify and genotype than SNVs. On the one hand, the long read lengths of cutting-edge single-molecule sequencing technologies – in particular, Pacific Biosciences (15-20 kb at 99.95% accuracy; Olson et al. 2022) and Oxford Nanopore Technologies (10-100 kb at 99.26% accuracy according to the Q20+ Simplex Dataset Release) – facilitate the reliable discovery of SVs of different types and sizes; however, high costs and low throughput still prohibit the routine application of these technologies in many research areas. On the other hand, high-throughput short paired-end read sequencing (e.g., to 2 × 150 bp NovaSeq at 99.92% accuracy; Olson et al. 2022) tends to be more affordable but SV detection can be hampered by high false discovery rates, particularly in repetitive, complex, and highly polymorphic regions of the genome which are prone to errors in base calling and alignment from short-read data (Cameron et al. 2019; Kosugi et al. 2019; and see the discussion in Mahmoud et al. 2019).

Together with the progress in sequencing technology, a variety of short-read whole-genome callers have been developed that utilize different signals in the sequencing data to computationally detect SVs (see Supplementary Table S1 of Kosugi et al. 2019 for a summary of popular short-read SV callers). Typically, in assembly-free approaches, these signals include regional differences in read depth, changes in the direction and/or distance between read pairs (i.e., discordant read pairs), as well as

unmatched read pairs that span SV breakpoints (i.e., split reads) (see Figure 2 in Alkan et al. 2011). Comprehensive benchmarking studies based on high-quality SV call sets obtained from deep-sequencing of human cell lines with multiple platforms (i.e., the *de facto* gold standard in the field) as well as simulated (ground truth) data have demonstrated that caller performance depends strongly on the SV type and size; as such, performance varies widely between approaches, with callers utilizing a combination of disparate read signals generally outperforming single-signal callers in terms of sensitivity (Cameron et al. 2017; 2019; Kosugi et al. 2019; Gabrielaite et al. 2021). For example, one of the largest benchmarking studies to date (Kosugi et al. 2019) showed that three of the best-performing short-read whole-genome SV callers – DELLY (Rausch et al. 2012), Lumpy (Layer et al. 2014), and Manta (Chen et al. 2016) – differ markedly in their precision and recall depending on the SV category. Specifically, although Lumpy performed best for very small (50 - 100 bp) and small (100 bp - 1 kb) deletions (with mean precision / recall rates for whole-genome human resequencing data ranging from 76.9% / 13.1% to 90.3% / 26.2%) as well as inversions (40.2% / 0.7%), Manta exhibited a higher precision for medium-sized (1 - 100 kb) and large (100 kb - 1 Mb) deletions (93.3% / 27.0% and 36.9% / 8.3%) as well as small and medium-sized duplications (54.9% / 7.4% and 19.0% / 0.9%). In contrast, DELLY outperformed both Lumpy and Manta in the detection of large duplications (7.0% / 2.4%). Given the complementary strengths of different methodologies, the authors also explored multi-caller scenarios, demonstrating that precision for different SV types and sizes can be improved by applying a so-called "ensemble" approach that generates a call set based on SVs detected by several independent callers (see Supplementary

Table S16 in Kosugi et al. 2019). This strategy is now widely employed in the field – though it should be noted that, unlike for SNVs (Pfeifer 2017b), no standardized best practices yet exist for SV discovery (see the discussion in Ho et al. 2020).

Methodologies aside, the great majority of work on the topic within primates to date has focused upon the great apes (Mao et al. 2024). However, in order to gain a broader evolutionary perspective, there would be great value in studying additional species from across the primate clade. One important evolutionary outgroup to the Haplorhini (which includes the apes), is the Daubentoniidae family of the Strepsirrhini suborder, which consists of the extinct giant aye-aye (*Daubentonia robusta*) (Nowak 1999) as well as extant aye-ayes (*Daubentonia madagascariensis*). Endemic to Madagascar, the world's largest nocturnal primate (Kay and Kirk 2000) inhabits primary rainforests and dry undergrowth forest on the eastern, northern, and north-western parts of the island (Sterling 1994a; Louis et al. 2020) – however, widespread forest degradation, fragmentation, as well as slash-and-burn agriculture continues to destroy many of their native habitats (Suzzi-Simmons 2023), posing a severe threat to the survival of the species. Aye-ayes exhibit many distinct phenotypic traits (Sterling and McCreless 2006), including elongated, flexible middle fingers and rodent-like teeth that allow them to extract small insects from decaying wood (Erickson 1994). In addition to wood-boring insects, their diet includes a variety of seeds and fruits, making them important seed dispensers in their native forests (Sterling 1994b). Consequently, aye-ayes play a crucial role in maintaining the general health and balance of Madagascar's flora and fauna. Yet, despite the aye-aye's unique ecological and evolutionary significance, our knowledge of the population genetics of this elusive species remains

limited (though see Perry et al. 2012, 2013 for insights into SNV diversity based on low-coverage sequencing data as well as Terbot et al. 2024). As one of the most critically endangered primate species on Earth (Louis et al. 2020), gaining insights into the structural variation landscape as a significant source of genetic diversity is thus vitally important for both conservation efforts of the species specifically, as well as to improve our understanding of the evolutionary history of primates in general.

Utilizing a high-precision ensemble approach for reliable SV discovery and genotyping from short-read sequencing data as described above, combined with previously developed methodology for SV curation in non-model organisms, we here analyze novel high-coverage genomic data of 14 individuals from ten trios (parents and their offspring) together with the recently released (Versoza and Pfeifer 2024), highly contiguous, well-annotated genome assembly for the species (a prerequisite for accurate SV discovery), to provide first insights into the genome-wide landscape of structural variation in this highly endangered primate.

RESULTS AND DISCUSSION

SVs were detected by whole-genome sequencing of 14 aye-aye individuals from ten parent-offspring trios (Supplementary Figure 1) to an average autosomal coverage of 46.1x (range: 41.8x to 49.0x; Supplementary Table 1) – well above the 30x generally recommended for SV discovery and genotyping from short-read data (see Wold et al. 2021 and references therein). In brief, as SV detection can be hampered by high

sequencing error rates and non-biological artefacts, raw reads were adapter and quality trimmed, before mapping them to the long-read genome assembly for the species and marking duplicates, which can result in spurious regions of extreme coverage (Supplementary Table 2). Based on these high-quality read mappings, SVs were then identified using an ensemble strategy that combined the strengths of local *de novo* assembly, with read depth, split-read, and discordant read approaches implemented in DELLY (Rausch et al. 2012), Lumpy (Layer et al. 2014), and Manta (Chen et al. 2016) – a methodology recently shown to result in robust and highly precise SV detection in humans (Subramanian et al. 2024). To increase precision, single-caller datasets were consolidated into a consensus call set of SVs identified by at least two of the three approaches using SURVIVOR (Jeffares et al. 2017) and subsequently filtered following the methodology described by Thomas et al. (2021) for another non-human primate species for which structural variation has been studied at the population-scale (rhesus macaque). In order to understand the potential medically-related impact of SVs, they were annotated using SnpEff (Cingolani et al. 2012), together with the gene annotations available from the aye-aye genome assembly (Versoza and Pfeifer 2024), and the putative relationship between large-effect SVs overlapping coding regions and diseases was assessed using the human database of Disease-Gene Associations with annotated Relationships among genes (eDGAR; Babbi et al. 2017) as a proxy.

A total of 1,133 autosomal SVs were identified in the 14 individuals, affecting 241,177 bp of the aye-aye genome (Figure 1). Of these 1,133 SVs, 1,000 were deletions (88.3%), 81 duplications (7.2%), 51 inversions (4.5%), and a single insertion – similar in proportion to the SV types previously observed in humans (89.4% deletions

and 10.6% duplications; Brandler et al. 2016; $\chi 2$ = 0.009; df = 1, *p*-value = 0.9226) and rhesus macaques (88.3% deletions and 11.7% duplications; Thomas et al. 2021; $\chi 2$ = 0.016; df = 1, *p*-value = 0.8993). In concordance with these earlier studies (Brasó-Vives et al. 2020; Thomas et al. 2021), the majority of segregating deletions were short (median length: 172 bp; Figure 2a) – a pattern presumably resulting from the fact that deletions are often associated with deleterious fitness effects and are thus purged from the population, with purifying selection having been observed to be acting more strongly on longer deletions which more easily perturb protein function (Taylor et al. 2004; Itsara et al. 2010; Mills et al. 2011; Yang et al. 2024). Duplications and inversions tended to be longer (median duplication / inversion length: 424 bp / 1.1 kb; Figure 2a); thus, while being smaller in number, each event affected a larger proportion of base-pairs on average (Figure 2b). From a technical standpoint, this pattern reflects, at least in part, an ascertainment bias as duplications and insertions are more difficult to detect from short-read sequencing data than deletions (see discussions in Conrad and Hurles 2007; Sudmant et al. 2015b; Kosugi et al. 2019; Mahmoud et al. 2019; Delage et al. 2020).

Per aye-aye individual, between 360 and 523 SVs were discovered (Supplementary Table 3) – a lower SV diversity than those previously observed in humans (Sudmant et al. 2015b) and rhesus macaques (Brasó-Vives et al. 2020; Thomas et al. 2021), consistent with the lower SNV diversity and effective population size of the species (Perry et al. 2012, 2013; Terbot et al. 2024). SVs were relatively evenly distributed across the genome (Supplementary Figure 2), with 12 SV-dense regions (≥ 10 SVs within a 10 Mb window) across the autosomal scaffolds (Supplementary Table 4). In concordance with observations in other primates (Bailey

and Eichler 2006; Brasó-Vives et al. 2020; Thomas et al. 2021), these SV-dense regions were enriched in sub-telomeric parts of the genome which frequently harbor transposable elements that facilitate non-allelic homologous recombination – a biological process mediating structural variation (Conrad and Hurles 2007). Moreover, the number of SVs was strongly correlated with the length of the scaffold (deletion: r = 0.977, p-value = 2.24 x 10^{-9}; duplications: r = 0.740, p-value = 0.0025; inversions: r = 0.798, p-value = 6.27 x 10^{-4}; Supplementary Figure 3) as previously observed in other organisms (Thomas et al. 2021; Wold et al. 2022). In agreement with previous work in humans (Conrad et al. 2010; Belyeu et al. 2021), SVs were frequently harbored in gene-rich regions, with 36.8% and 20.4% residing within intronic and non-exonic, non-frameshift, non-missense genic regions, respectively. In addition to SVs within intergenic regions (41.6%), six SVs caused frameshifts, four SVs were located in exonic regions, three SVs were stop-related, and one SV resulted in a missense mutation (Supplementary Table 5). A total of 625 SVs (55.2%) were predicted to affect transcripts and a further 35 SVs (3.1%) were putative gene variants; the remainder were predicted to impact intergenic features. The vast majority of SVs were classified as modifiers (94.7%); the remaining SVs were predicted to have a high (4.0%), moderate (0.9%), and low (0.4%) impact (Figure 3), including several potential gene fusion events, exon losses, and frameshift mutations (Table 1). As expected, SVs with predicted high, moderate, and low effects were significantly enriched in genic regions ($\chi 2$ = 40.902; df = 1, p-value = 1.6 $\times$ 10^{-10}). Out of the 45 major effect SVs, two deletions were located within disease-linked genes: (i) a frameshift variant in the cleavage factor polyribonucleotide kinase subunit 1 (CLP1) gene linked to pontocerebellar hypoplasia

(PCH) subtype 10 – an autosomal recessive condition characterized by impaired brain development, motor neuron degeneration, and seizures (Karaca et al. 2014; Schaffer et al. 2014; and see review by van Dijk et al. 2018) – and (ii) a variant leading to a gene fusion in the opioid binding protein/cell adhesion molecule like (OPCML) gene – a tumor suppressor that is often epigenetically silenced in cancer, most prominently ovarian cancer (Birtley et al. 2019) (Table 1). Although the remaining major effect SVs were not predicted to exhibit a direct link to a disease, several ablated or disrupted genes, including those related to immune response (IGHV1-18 and IGHV1-24; Rodriguez et al. 2023) as well as circadian rhythm, particularly diurnal oscillations in light and temperature (BMAL2; Pando et al. 2001), were observed. Furthermore, several major effect SVs were located outside of coding regions and future work focusing on the potential regulatory impact of these changes would thus be of great interest.

The three-generation pedigree structure of this study also offered an opportunity to study Mendelian inheritance in the ten parent-offspring trios. Out of the 1,133 identified SVs, 114 sites (10.1%, including 99 deletions, eight duplications, and seven inversions) exhibited genotype-based Mendelian inconsistencies and were thus independently visualized for validation – a strategy previously shown to be in agreement with other orthogonal validation techniques such as ddPCR and long-read sequencing (Bertolotti et al. 2020; Belyeu et al. 2021). Perhaps unsurprisingly, given the orders of magnitude lower rates of structural mutation compared to point mutation previously observed in other primates (Werling et al. 2018), the lower genetic diversity of aye-ayes compared to the great apes (Perry et al. 2012, 2013; Terbot et al. 2024), and the small number of individuals in the cohort, no genuine *de novo* SVs were detected.

In addition to the small pedigree preventing the detection of rare SVs, a general caveat of short-read approaches, such as the ones employed here, is their inability to accurately identify SVs in low-complexity, highly repetitive regions of the genome (Chaisson et al. 2019). Due to the high false discovery rates frequently observed in such regions (Cameron et al. 2019; Kosugi et al. 2019; Mahmoud et al. 2019), regions harboring gaps, repeats, and/or extreme coverage were excluded from this study to increase precision. Additionally, stringent filtering was applied to the remaining regions to limit the dataset to SVs of high confidence and avoid spurious calls. It should be noted, however, that such an exclusion and filtering will necessarily lead to an underestimation of SVs, particularly those driven by homology-mediated mechanisms such as non-allelic homologous recombination, fork stalling and template switching, and microhomology-mediated break-induced replication (Belyeu et al. 2021).

In order to assess the SV calling and genotyping accuracy of the employed ensemble approach, the 1,596 genotypes at the 114 sites flagged as Mendelian violations were manually inspected in the pedigree. Interestingly, out of the curated sites, 35 (30.7%) were fixed for the reference allele (i.e., there was no read support for a SV at the site) and six (5.3%) were fixed for the alternative allele (Supplementary Table 6). Moreover, in contrast to a previous simulation study which observed high genotyping precision for deletions (91.9–94.1%), duplications (59.7–84.3%), inversions (87.8–88.0%), and insertions (97.7%) in all three callers (Supplementary Table S15 in Kosugi et al. 2019), visual curation of 1,596 genotypes at the sites of Mendelian violations in the pedigreed dataset revealed a high rate of genotyping error (35.7%; Supplementary Table 6, and see Supplementary Figure 4 for an example of an incorrectly called

genotype detected during the manual review). Taken together, these observations highlight that accurate SV calling and genotyping based on short-read data remains challenging (see also the discussion in Sibbesen et al. 2018), and emphasizes the importance of manual curation in studies of structural variation.

CONCLUSION

With fewer than 1,000 to 10,000 individuals estimated to remain in the wild, aye-ayes are imminently threatened by extinction. Gaining insights into the genetic diversity of the species is thus of vital importance, and the first view of the landscape of structural variation presented here will be crucial to advance our understanding of the connection between genotypes and phenotypic traits relevant to conservation efforts and species recovery. In addition, as an important outgroup to the Haplorhini, this genomic data will allow for deeper comparative analyses across the primate clade to further our understanding of primate evolutionary history. In this regard, it is important to keep in mind that, although many similarities emerged between the structural variant landscape of aye-ayes and those of other primates studied to date, SV discovery approaches can have large impacts on the accuracy of both SV calls and their genotypes, thus hindering quantitative comparisons across studies. Moving forward, to facilitate meaningful comparisons, an important emphasis of future comparative genomic studies should thus be on the development of streamlined, uniform pipelines across the primate clade. Moreover, as short-read approaches are biased with regards to the SV types and sizes

that they are able to detect, future studies should, whenever feasible, complement short-read data with long-read and/or optical sequencing approaches to obtain insights into the full spectrum of structural variation, including translocations and complex SVs. Ultimately, there is a pressing need to combine novel genomic resources, such as the one presented here, with ecological and evolutionary research in order to aid the development of more effective conservation strategies for this charismatic species.

MATERIALS AND METHODS

Animal subjects

This study was approved by the Duke Lemur Center's Research Committee (protocol BS-3-22-6) and Duke University's Institutional Animal Care and Use Committee (protocol A216-20-11). The study was performed in compliance with all regulations regarding the care and use of captive primates, including the U.S. National Research Council's Guide for the Care and Use of Laboratory Animals and the U.S. Public Health Service's Policy on Human Care and Use of Laboratory Animals.

Sample collection, preparation, and sequencing

Genomic DNA was extracted from peripheral blood samples of 14 aye-aye (*D. madagascariensis*) individuals (six males and eight females) originating from a single three-generation pedigree using the PureLink Genomic DNA Mini Kit (ThermoFisher Scientific, Waltham, MA, USA) and quantified using a Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, MA, USA). Following manufacturer's instructions, a sequencing library was prepared for each sample using the NEBNext® Ultra™ II DNA PCR-free Library Prep Kit (New England, Ipswich, MA, USA). Quality control of each library was performed using a High Sensitivity D1000 ScreenTape on an Agilent TapeStation (Agilent Technologies, Palo Alto, CA, USA). Libraries were quantified using a Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, MA, USA) and real-time PCR (Applied Biosystems, Carlsbad, CA, USA). Each library was paired-end

sequenced (2 × 150 bp) on an Illumina NovaSeq platform (Illumina, San Diego, CA, USA).

Read mapping

FastQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Cutadapt v.1.18 (https://cutadapt.readthedocs.io/en/stable/) embedded within TrimGalore v.0.6.10 (https://github.com/FelixKrueger/TrimGalore) were used to trim low-quality bases (with a Phred score < 20) and remove Illumina adapter sequences from the 3'-ends of the reads as they can lead to incorrect mappings. Afterward, the quality-controlled reads were mapped to the chromosome-level genome assembly for the species (DMad_hybrid; GenBank accession number: JBFSEQ000000000; Versoza and Pfeifer 2024) using BWA-MEM v.0.7.17 (Li and Durbin 2009). Read mappings were sorted, duplicates marked, and indexed using SAMtools *sort* v.1.20 (Danecek et al. 2021), GATK4 *MarkDuplicates* v.4.5 (Van der Auwera and O'Connor 2020), and SAMtools *index* v.1.20, respectively.

Quality control

SV detection can be hampered by high sequencing error rates, uneven read coverage, and/or skewed insert size distributions, thus the quality of the read mappings and coverage distributions for each individual were assessed using SAMtools v.1.16 (Danecek et al. 2021) and goleft v.0.2.6 (https://github.com/brentp/goleft) prior to variant calling. Moreover, as regions harboring gaps, repeats, and/or extreme coverage frequently lead to mapping errors (Mahmoud et al. 2019), such genomic regions were

excluded during the variant calling. In brief, sample coverage was estimated with mosdepth v.0.3.8 (Pedersen and Quinlan 2018) and high-coverage regions (defined here as regions exhibiting more than 10-fold of the mean autosomal coverage) as well as repetitive regions (including retroelements, DNA transposons, simple repeats, and low-complexity repeats) annotated in the aye-aye genome assembly (Versoza and Pfeifer 2024) were excluded.

SV calling and genotyping

To increase precision, autosomal SVs were jointly called in the 14 aye-aye individuals using three of the best-performing short-read whole-genome SV callers according to recent benchmarking studies (Kosugi et al. 2019; Gabrielaite et al. 2021): DELLY v.1.2.6 (Rausch et al. 2012), Manta v.1.6.0 (Chen et al. 2016) and Lumpy v.0.2.13 (Layer et al. 2014) embedded within Smoove v.0.2.6 (https://github.com/brentp/smoove).

DELLY uses a combination of paired-end, read depth, and split-read signals for SV discovery. DELLY *call* was used to detect SVs from the read mappings, excluding (*-x*) repetitive and high-coverage regions as detailed above. Low-quality (*LowQual*) calls with fewer than three paired-end (*PE*) reads supporting a variant or with a mean mapping quality of less than 20 were discarded using BCFtools *view* v.1.10.2 with the *-e ' FILTER=="LowQual" || FORMAT/FT=="LowQual" '* flag, limiting the call set to precise SVs with split-read support at nucleotide resolution.

Manta combines paired- and split-read signals to detect, assemble, and genotype SVs. A configuration file was created (using the built-in *configManta.py* script)

that provides information on the samples (--*bam*) and reference assembly (--*referenceFasta*) before running the two-step workflow (*runWorkflow.py*), consisting of a genome scan to identify candidate regions, followed by SV discovery, breakend assembly, genotyping, and filtering. Reported inversions were reformatted into single inverted sequence junctions using the built-in *convertInversions.py* script. SV calls were limited to variants outside of repetitive and high-coverage regions that passed all filter criteria using VCFtools --*exclude-bed* v.0.1.14 (Danecek et al. 2011) and BCFtools *view* v.1.10.2 with the *-i ' FILTER=="PASS" '* option, respectively. In brief, these filters excluded low-quality SVs (QUAL < 20 and, for those smaller than 1kb, sites where the proportion of reads in all individuals with a MAPQ0 around the breakend exceeds 40%), SVs larger than the paired-end fragment size without paired-end read support for the alternative allele, deletions and duplications inconsistent with diploid expectations, as well as SVs with breakends occurring in regions of excessive read depth (defined here as more than three times the median chromosome depth).

Lumpy utilizes regional differences in read depth to identify SVs; in addition, Lumpy detects unmatched read pairs by extracting split-read alignments (using the built-in *extractSplitReads_BwaMem* script) from discordant paired-end alignments (obtained using the ' *samtools view -b -F 1294* ' command). To accelerate the Lumpy workflow, the Smoove wrapper script *call* was used to parallelize these different steps, calling SVs outside of problematic regions (--*exclude*) and directly genotyping (--*genotype*) detected SVs using the Bayesian likelihood genotyper SVTyper v.0.7.0 (Chiang et al. 2015). By default, Smoove implements a series of filters that remove spurious alignments and improve specificity. Specifically, Smoove excluded reads that

were soft-clipped at both ends, contained more than three mismatches, or exhibited alternative matches. To avoid spurious calls, Smoove further discarded split-reads for which the reads in a pair mapped to different chromosomes, split or discordant reads with a high depth of coverage (> 1000) as well as orphaned reads (i.e., reads without a mate). Following the developer's recommendations (https://github.com/brentp/smoove), calls were annotated using *smoove annotate* and limited to sites with high-quality heterozygotes (i.e., SVs with a mean Smoove heterozygote quality [MSHQ] score larger than 3). Additionally, deletions and duplications were limited to sites with a fold-change of variant depth relative to flanking regions (DHFFC) of less than 0.7 and relative to genomic regions with similar GC-content (DHBFC) larger than 1.3, respectively.

In order to obtain high-precision calls, individual, single-caller datasets were consolidated into a consensus call set of SVs identified by at least two of the three approaches using SURVIVOR *merge* v.1.0.7 (Jeffares et al. 2017), merging any SVs of the same type that are closer than 500 bp.

SV filtering

To reduce false positives, the consensus call set was filtered following the methodology described by Thomas et al. (2021). In brief, SVs present in all or all but one individual were removed as these are likely the result of local mis-assembly. Furthermore, SVs larger than 100 kb as well as those of low quality (QUAL < 100) were excluded to further limit the number of spurious variants in the dataset. The remaining SVs were then annotated with read depth information using Duphold v.0.2.1 (Pedersen

and Quinlan 2019) embedded within Smoove v.0.2.6, and deletion and duplication events were limited to those exhibiting a fold-change of coverage of <0.7 and >1.3, respectively. Lastly, putative *Alu* mobile element insertions were filtered out by removing any SVs with a length between 275 and 325 bp.

Functional annotation

SVs were annotated using SnpEff v.5.2 (Cingolani et al. 2012) based on gene annotations available in the aye-aye genome assembly (Versoza and Pfeifer 2024). In order to understand the potential medically-related impact of SVs, the putative relationship between large-effect SVs overlapping coding regions and diseases was assessed using the database of Disease-Gene Associations with annotated Relationships among genes (eDGAR; Babbi et al. 2017), with information from the Online Mendelian Inheritance in Man (OMIM; Amberger et al. 2017), humsavar (UniProt Consortium et al. 2015), and ClinVar (Landrum et al. 2016) databases embedded within.

Identification of *de novo* SVs and assessment of SV calling / genotyping accuracy

Based on the final SV call set, Mendelian violations were identified using BCFtools v.1.20 (Danecek et al. 2021) with the +*mendelian* plugin and visually reviewed using Samplot v.1.3.0 (Belyeu et al. 2021) to identify *de novo* SVs and assess SV calling / genotyping accuracy.

ACKNOWLEDGEMENTS

We would like to thank Erin Ehmke, Kay Welser, and the Duke Lemur Center for providing the aye-aye samples used in this study, and Fritz Sedlazeck as well as members of the Jensen Lab and Pfeifer Lab for helpful discussions. DNA extraction, library preparation, and Illumina sequencing was conducted at Azenta Life Sciences (South Plainfield, NJ, USA). Computations were performed on the Sol supercomputer at Arizona State University (Jennewein et al. 2023). This is Duke Lemur Center publication # XXXX.

FUNDING

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM151008 to SPP and the National Science Foundation under Award Number DBI-2012668 to the Duke Lemur Center. CJV was supported by the National Science Foundation CAREER Award DEB-2045343 to SPP. JDJ was supported by National Institutes of Health Award Number R35GM139383. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

REFERENCES

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 12(5):363–376.

Amberger JS, Hamosh A. 2017. Searching Online Mendelian Inheritance in Man (OMIM): a knowledgebase of human genes and genetic phenotypes. *Curr Protoc Bioinformatics*. 58:1.2.1–1.2.12.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J, et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science*. 336(6078):193–198.

Babbi G, Martelli PL, Profiti G, Bovo S, Savojardo C, Casadio R. 2017. eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genomics*. 18(Suppl 5):554.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*. 7(7):552–564.

Belyeu JR, Chowdhury M, Brown J, Pedersen BS, Cormier MJ, Quinlan AR, Layer RM. 2021. Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol*. 22(1):161.

Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, Robledo D, Kent MP, Røsæg LL, Holen MM, et al. 2020. The structural variation landscape in 492 Atlantic salmon genomes. *Nat Commun*. 11(1):5176.

Birtley JR, Alomary M, Zanini E, Antony J, Maben Z, Weaver GC, Von Arx C, Mura M, Marinho AT, Lu H, et al. 2019. Inactivating mutations and X-ray crystal structure of the tumor suppressor OPCML reveal cancer-associated functions. *Nat Commun*. 10(1):3134.

Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, Barrera DJ, Lin GN, Malhotra D, Watts AC, et al. 2016. Frequency and complexity of *de novo* structural mutation in autism. *Am J Hum Genet*. 98(4):667–679.

Brasó-Vives M, Povolotskaya IS, Hartasánchez DA, Farré X, Fernandez-Callejo M, Raveendran M, Harris RA, Rosene DL, Lorente-Galdos B, Navarro A, et al. 2020. Copy number variants and fixed duplications among 198 rhesus macaques (*Macaca mulatta*). *PLoS Genet*. 16(5):e1008742.

Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun*. 10(1):3240.

Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT. 2017. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res*. 27(12):2050–2060.

Chaignat E, Yahya-Graison EA, Henrichsen CN, Chrast J, Schütz F, Pradervand S, Reymond A. 2011. Copy number variation modifies expression time courses. *Genome Res*. 21(1):106–113.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 10(1):1784.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 32(8):1220–1222.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*. 12(10):966–968.

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEx Consortium, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet*. 49(5):692–699.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 6(2):80–92.

Conrad DF, Hurles ME. 2007. The population genetics of structural variation. *Nat Genet*. 39(7 Suppl):S30–S36.

Conrad DF, Pinto D, Redon R, Feuk L, Gökçümen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature*. 464(7289):704–712.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics*. 27(15):2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Giga Science*. 10(2):giab008.

Delage WJ, Thevenon J, Lemaitre C. 2020. Towards a better understanding of the low recall of insertion variants with short-read based variant callers. *BMC Genomics*. 21(1):762.

Erickson CJ. 1994. Tap-scanning and extractive foraging in aye-ayes, *Daubentonia madagascariensis*. *Folia Primatol (Basel)*. 62(1-3):125–135.

Gabrielaite M, Torp MH, Rasmussen MS, Andreu-Sánchez S, Vieira FG, Pedersen CB, Kinalis S, Madsen MB, Kodama M, Demircan GS, et al. 2021. A comparison of

tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers (Basel)*. 13(24):6283.

Ghafoor S, Santos J, Versoza CJ, Jensen JD, Pfeifer SP. 2023. The impact of sample size and population history on observed mutational spectra: a case study in human and chimpanzee populations. *Genome Biol Evol*. 15(3):evad019.

Girirajan S, Campbell CD, Eichler EE. 2011. Human copy number variation and complex genetic disease. *Annu Rev Genet*. 45:203–226.

Gökçümen O, Lee C. 2009. Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. *Methods*. 49(1):18–25.

Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nat Rev Genet*. 21(3):171–189.

Hollox EJ, Zuccherato LW, Tucci S. 2022. Genome structural variation in human evolution. *Trends Genet*. 38(1):45–58.

Iskow RC, Gökçümen O, Lee C. 2012. Exploring the role of copy number variants in human adaptation. *Trends Genet*. 28(6):245–257.

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. 2010. *De novo* rates and selection of large copy number variation. *Genome Res*. 20(11):1469–1481.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun*. 8:14061.

Jennewein DM, Lee J, Kurtz C, Dizon W, Shaeffer I, Chapman A, Chiquete A, Burks J, Carlson A, Mason N, et al. 2023. The Sol Supercomputer at Arizona State University. In Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good (PEARC '23). Association for Computing Machinery, New York, NY, USA, 296–301.

Johri P, Pfeifer SP, Jensen JD. 2023. Developing an evolutionary baseline model for humans: jointly inferring purifying selection with population history. *Mol Biol Evol*. 40(5):msad100.

Karaca E, Weitzer S, Pehlivan D, Shiraishi H, Gogakos T, Hanada T, Jhangiani SN, Wiszniewski W, Withers M, Campbell IM, et al. 2014. Human CLP1 mutations alter tRNA biogenesis, affecting both peripheral and central nervous system function. *Cell*. 157(3):636–650.

Kay RF, Kirk EC. 2000. Osteological evidence for the evolution of activity pattern and visual acuity in primates. *Am J Phys Anthropol*. 113(2):235–262.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 20(1):117.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 44(D1):D862–D868.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 15(6):R84.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25(14):1754–1760.

Lin YL, Gökçümen O. 2019. Fine-scale characterization of genomic structural variation in the human genome reveals adaptive and biomedically relevant hotspots. *Genome Biol Evol*. 11(4):1136–1151.

Louis EE, Sefczek TM, Randimbiharinirina DR, Raharivololona B, Rakotondrazandry JN, Manjary D, Aylward M, Ravelomandrato F. 2020. *Daubentonia madagascariensis*. The IUCN red list of threatened species. Version 2020.2; e.T6302A115560793 doi:10.2305/IUCN.UK.2020-2.RLTS.T6302A115560793.en.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol*. 20(1):246.

Mao Y, Harvey WT, Porubsky D, Munson KM, Hoekzema K, Lewis AP, Audano PA, Rozanski A, Yang X, Zhang S, et al. 2024. Structurally divergent and recurrently mutated regions of primate genomes. *Cell*. 187(6):1547–1562.e13.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, et al. 2011. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*. 21(6):830–839.

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017. Tracing the peopling of the world through genomics. *Nature*. 541(7637):302–310.

Nowak, RM. 1999. *Walker's Primates of the World* (6th ed.). 1999. Baltimore, Maryland: Johns Hopkins University Press. ISBN 978-0-8018-6251-9.

Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, et al. 2022. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom*. 2(5):100129.

Pando MP, Pinchak AB, Cermakian N, Sassone-Corsi P. 2001. A cell-based system that recapitulates the dynamic light-dependent regulation of the vertebrate clock. *Proc Natl Acad Sci U S A*. 98(18):10178–10183.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol*. 11(5):R52.

Pedersen BS, Quinlan AR. 2019. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *Giga Science*. 8(4):giz040.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 34(5):867–868.

Perry GH, Louis EE Jr, Ratan A, Bedoya-Reina OC, Burhans RC, Lei R, Johnson SE, Schuster SC, Miller W. 2013. Aye-aye population genomic analyses highlight an important center of endemism in northern Madagascar. *Proc Natl Acad Sci U S A*. 110(15):5823–5828.

Perry GH, Melsted P, Marioni JC, Wang Y, Bainer R, Pickrell JK, Michelini K, Zehr S, Yoder AD, Stephens M, et al. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res*. 22(4):602–610.

Pfeifer SP. 2020a. A fine-scale genetic map for vervet monkeys. *Mol Biol Evol*. 37(7):1855–1865.

Pfeifer SP. 2017a. Direct estimate of the spontaneous germ line mutation rate in African green monkeys. *Evolution*. 71(12):2858–2870.

Pfeifer SP. 2017b. From next-generation resequencing reads to a high-quality variant data set. *Heredity (Edinb)*. 118(2):111–124.

Pfeifer SP. 2020b. Spontaneous mutation rates. In The Molecular Evolutionary Clock. Theory and Practice. Springer Nature.

Pfeifer SP. 2021. Studying mutation rate evolution in primates – the effects of computational pipelines and parameter choices. *Giga Science*. 10(10):giab069.

Pfeifer SP, Jensen JD. 2016. The impact of linked selection in chimpanzees: a comparative study. *Genome Biol Evol*. 8(10):3202–3208.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 28(18):i333–i339.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature*. 444(7118):444–454.

Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, Tieri D, Ke H, Jackson KJL, Boyd SD, et al. 2023. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat Commun*. 14(1):4419.

Schaffer AE, Eggens VR, Caglayan AO, Reuter MS, Scott E, Coufal NG, Silhavy JL, Xue Y, Kayserili H, Yasuno K, et al. 2014. CLP1 founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. *Cell*. 157(3):651–663.

Sibbesen JA, Maretty L, Danish Pan-Genome Consortium, Krogh A. 2018. Accurate genotyping across variant classes and lengths using variant graphs. *Nat Genet*. 50(7):1054–1059.

Simkin AT, Bailey JA, Gao FB, Jensen JD. 2014. Inferring the evolutionary history of primate microRNA binding sites: overcoming motif counting biases. *Mol Biol Evol*. 31(7):1894–1901.

Soni V, Jensen JD. 2024. Inferring demographic and selective histories from population genomic data using a two-step approach in species with coding-sparse genomes: an application to human data. BioRxiv, preprint.

Soni V, Pfeifer SP, Jensen JD. 2024a. The effects of mutation and recombination rate heterogeneity on the inference of demography and the distribution of fitness effects. *Genome Biol Evol*. 16(2):evae004.

Soni V, Terbot JW, Versoza CJ, Pfeifer SP, Jensen JD. 2024b. A whole-genome scan for evidence of recent positive and balancing selection in aye-ayes (*Daubentonia madagascariensis*) utilizing a well-fit evolutionary baseline model. BioRxiv, preprint.

Sterling E. 1994a. Taxonomy and distribution of Daubentonia: a historical perspective. *Folia Primatol (Basel)*. 62(1-3):8–13.

Sterling EJ. 1994b. Aye-ayes: specialists on structurally defended resources. *Folia Primatol (Basel)*. 62(1-3):142–154.

Sterling EJ, McCreless E. 2006. Adaptations in the aye-aye: a review. In: Gould L, Sauther M, editors. Lemurs: ecology and adaptation. New York (NY): Springer. p. 159-184.

Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, Great Ape Genome Project, Bustamante CD, Hammer MF, Wall JD. 2016. The time scale of recombination rate evolution in great apes. *Mol Biol Evol*. 33(4):928–945.

Subramanian K, Chopra M, Kahali B. 2024. Landscape of genomic structural variations in Indian population-based cohorts: deeper insights into their prevalence and clinical relevance. *HGG Adv*. 5(3):100285.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res*. 23(9):1373–1382.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science*. 330(6004):641–646.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015a. Global diversity, population stratification, and selection of human copy-number variation. *Science*. 349(6253):aab3761.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. 2015b. An integrated map of structural variation in 2,504 human genomes. *Nature*. 526(7571):75–81.

Suzzi-Simmons A. 2023. Status of deforestation of Madagascar. *Glob Ecol Conserv*. 42:e02389.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res*. 14(4):555–566.

Terbot JW, Soni V, Versoza CJ, Pfeifer SP, Jensen JD. 2024. Inferring the demographic history of aye-ayes (*Daubentonia madagascariensis*) from high-quality, whole-genome, population-level data. BioRxiv, preprint.

Thomas GWC, Wang RJ, Nguyen J, Harris RA, Raveendran M, Rogers J, Hahn MW. 2021. Origins and long-term patterns of copy-number variation in rhesus macaques. *Mol Biol Evol*. 38(4):1460–1471.

UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Res*. 43(Database issue):D204–D212.

van der Auwera GA, O'Connor BD. 2020. Genomics in the cloud: using Docker, GATK, and WDL in Terra. Sebastopol: O'Reilly Media.

van Dijk T, Baas F, Barth PG, Poll-The BT. 2018. What's new in pontocerebellar hypoplasia? An update on genes and subtypes. *Orphanet J Rare Dis*. 13(1):92.

Versoza CJ, Jensen JD, Pfeifer SP. 2024b. Characterizing the rates and patterns of *de novo* germline mutations in the aye-aye (*Daubentonia madagascariensis)*. BioRxiv, preprint.

Versoza CJ, Lloret-Villas A, Jensen JD, Pfeifer SP. 2024. A pedigree-based map of crossovers and non-crossovers in aye-ayes (*Daubentonia madagascariensis)*. BioRxiv, preprint.

Versoza CJ, Pfeifer SP. 2024. A hybrid genome assembly of the endangered aye-aye (*Daubentonia madagascariensis*). *G3 (Bethesda)*. 14(10): jkae185.

Versoza CJ, Weiss S, Johal R, La Rosa B, Jensen JD, Pfeifer SP. 2024a. Novel insights into the landscape of crossover and noncrossover events in rhesus macaques (*Macaca mulatta*). *Genome Biol Evol*. 16(1):evad223.

Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-Papadimitriou E, et al. 2018. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet*. 50(5):727–736.

Wold J, Koepfli KP, Galla SJ, Eccles D, Hogg CJ, Le Lec MF, Guhlin J, Santure AW, Steeves TE. 2021. Expanding the conservation genomics toolbox: incorporating structural variants to enhance genomic studies for species of conservation concern. *Mol Ecol*. 30(23):5949–5965.

Wold JR, Guhlin JG, Dearden PK, Santure AW, Steeves TE. 2023. The promise and challenges of characterizing genome-wide structural variants: a case study in a critically endangered parrot. *Mol Ecol Resour*. doi: 10.1111/1755-0998.13783. Epub ahead of print. PMID: 36916824.

Yang Y, Braga MV, Dean MD. 2024. Insertion-deletion events are depleted in protein regions with predicted secondary structure. *Genome Biol Evol*. 16(5):evae093.

Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. *Nat Rev Genet*. 16(3):172–183.

Table 1. **Structural variants with major effects in aye-ayes**. SVs in disease-linked genes are highlighted in red.

scaffold	start	size	type	predicted effect	allele freq.	feature type	transcript type	eDGAR
scaffold 1	114,706,693	-1,678	DEL	splice donor variant	0.18	transcript	pseudogene	N/A
scaffold 1	117,064,874	65,741	DUP	stop gained	0.04	transcript	protein-coding	not a disease-linked gene
scaffold 1	117,231,258	-3,148	DEL	gene fusion	0.46	gene variant	–	not a disease-linked gene
scaffold 1	175,090,637	-3,775	DEL	gene fusion	0.14	gene variant	–	not directly linked to a disease
scaffold 1	206,727,921	-888	DEL	frameshift variant; splice acceptor & donor variant	0.43	transcript	protein-coding	disease-linked gene
scaffold 1	308,586,685	-2,940	DEL	gene fusion	0.14	gene variant	–	disease-linked gene
scaffold 2	193,565	65,573	DUP	gene fusion	0.36	gene variant	–	not a disease-linked gene
scaffold 2	16,532,751	-1,563	DEL	splice acceptor variant	0.32	transcript	pseudogene	N/A
scaffold 2	59,474,585	-1,613	DEL	splice donor variant	0.25	transcript	pseudogene	N/A
scaffold 2	123,894,446	-265	DEL	splice donor variant	0.43	transcript	pseudogene	N/A
scaffold 2	147,696,112	-1,992	DEL	splice donor variant	0.18	transcript	pseudogene	N/A
scaffold 2	167,674,811	-4,633	DEL	transcript ablation	0.36	transcript	protein-coding	not directly linked to a disease
scaffold 2	280,664,663	-352	DEL	frameshift variant; splice acceptor & donor variant	0.25	transcript	protein-coding	not a disease-linked gene
scaffold 2	280,679,454	-607	DEL	splice acceptor & donor variant	0.21	transcript	protein-coding	not a disease-linked gene
scaffold 3	62,048,361	-56	DEL	frameshift variant	0.32	transcript	protein-coding	not a disease-linked gene
scaffold 3	114,478,661	24,999	DUP	splice donor variant	0.11	transcript	pseudogene	not directly linked to a disease
scaffold 3	161,585,894	-949	DEL	splice acceptor variant	0.54	transcript	pseudogene	not directly linked to a disease
scaffold 3	195,072,212	-1,137	DEL	exon loss variant	0.18	transcript	pseudogene	not directly linked to a disease
scaffold 3	224,565,917	-168	DEL	splice acceptor variant	0.18	transcript	protein-coding	not directly linked to a disease
scaffold 3	226,807,718	-93	DEL	frameshift variant; splice donor variant	0.04	transcript	protein-coding	not a disease-linked gene
scaffold 3	240,213,435	-54	DEL	splice donor variant	0.11	transcript	protein-coding	not a disease-linked gene
scaffold 4	210,144,666	-53,110	DEL	splice acceptor & donor variant	0.21	transcript	protein-coding	not a disease-linked gene
scaffold 5	14,376,755	-93	DEL	splice donor variant	0.18	transcript	pseudogene	not directly linked to a disease
scaffold 5	18,775,900	-3,663	DEL	exon loss variant; splice acceptor & donor variant	0.36	transcript	protein-coding	not a disease-linked gene
scaffold 5	72,903,924	2,895	DUP	stop gained	0.39	transcript	protein-coding	not directly linked to a disease
scaffold 5	80,139,609	-167	DEL	exon loss variant; stop lost; splice acceptor & donor variant	0.18	transcript	protein-coding	not directly linked to a disease
scaffold 5	94,879,702	-34,807	DEL	feature ablation	0.54	gene variant	–	not a disease-linked gene
scaffold 5	167,714,291	-1,391	DEL	gene fusion	0.14	gene variant	–	not a disease-linked gene
scaffold 5	170,514,034	480	DUP	frameshift variant; splice acceptor & donor variant	0.14	transcript	protein-coding	not a disease-linked gene
scaffold 5	191,966,847	-801	DEL	splice donor variant	0.07	transcript	protein-coding	not a disease-linked gene
scaffold 6	146,740,803	-66	DEL	splice donor variant	0.25	transcript	pseudogene	not directly linked to a disease
scaffold 7	27,195,024	-1,624	DEL	gene fusion	0.61	gene variant	–	not a disease-linked gene
scaffold 7	40,345,354	-1,612	DEL	bidirectional gene fusion	0.04	gene variant	–	not directly linked to a disease
scaffold 7	164,612,242	-210	DEL	splice acceptor & donor variant	0.21	transcript	protein-coding	not a disease-linked gene
scaffold 7	188,327,797	-58	DEL	splice acceptor variant	0.39	transcript	pseudogene	not directly linked to a disease
scaffold 8	8,255,327	857	INV	bidirectional gene fusion	0.04	gene variant	–	not directly linked to a disease
scaffold 8	55,672,556	-1,536	DEL	frameshift variant; splice donor variant	0.21	transcript	protein-coding	not a disease-linked gene
scaffold 8	55,677,369	-163	DEL	splice acceptor variant	0.21	transcript	protein-coding	not a disease-linked gene
scaffold 8	55,677,669	-2,837	DEL	exon loss variant; splice acceptor & donor variant	0.21	transcript	protein-coding	not a disease-linked gene
scaffold 10	84,664,633	-90	DEL	stop lost	0.11	transcript	protein-coding	not a disease-linked gene
scaffold 11	39,728,906	-1,012	DEL	splice acceptor & donor variant	0.25	transcript	protein-coding	not a disease-linked gene
scaffold 12	52,387,052	-5,359	DEL	splice acceptor & donor variant	0.07	transcript	pseudogene	not directly linked to a disease
scaffold 13	26,376,628	-238	DEL	splice acceptor & donor variant	0.43	transcript	protein-coding	not a disease-linked gene
scaffold 13	26,376,969	-990	DEL	splice acceptor & donor variant	0.43	transcript	protein-coding	not a disease-linked gene
scaffold 13	51,852,118	-42,903	DEL	feature ablation	0.11	gene variant	–	not a disease-linked gene

Figure 1. **Landscape of structural variation in the aye-aye genome**. Genome-wide map of structural variation (deletions are color-coded in blue, duplications in teal, insertions in yellow, and inversions in olive green) across autosomal scaffolds (note that scaffold 9, i.e., chromosome X, is not displayed), with peak height being proportional to the SV length. Putative *Alu* elements (shown in red) were removed prior to analyses.

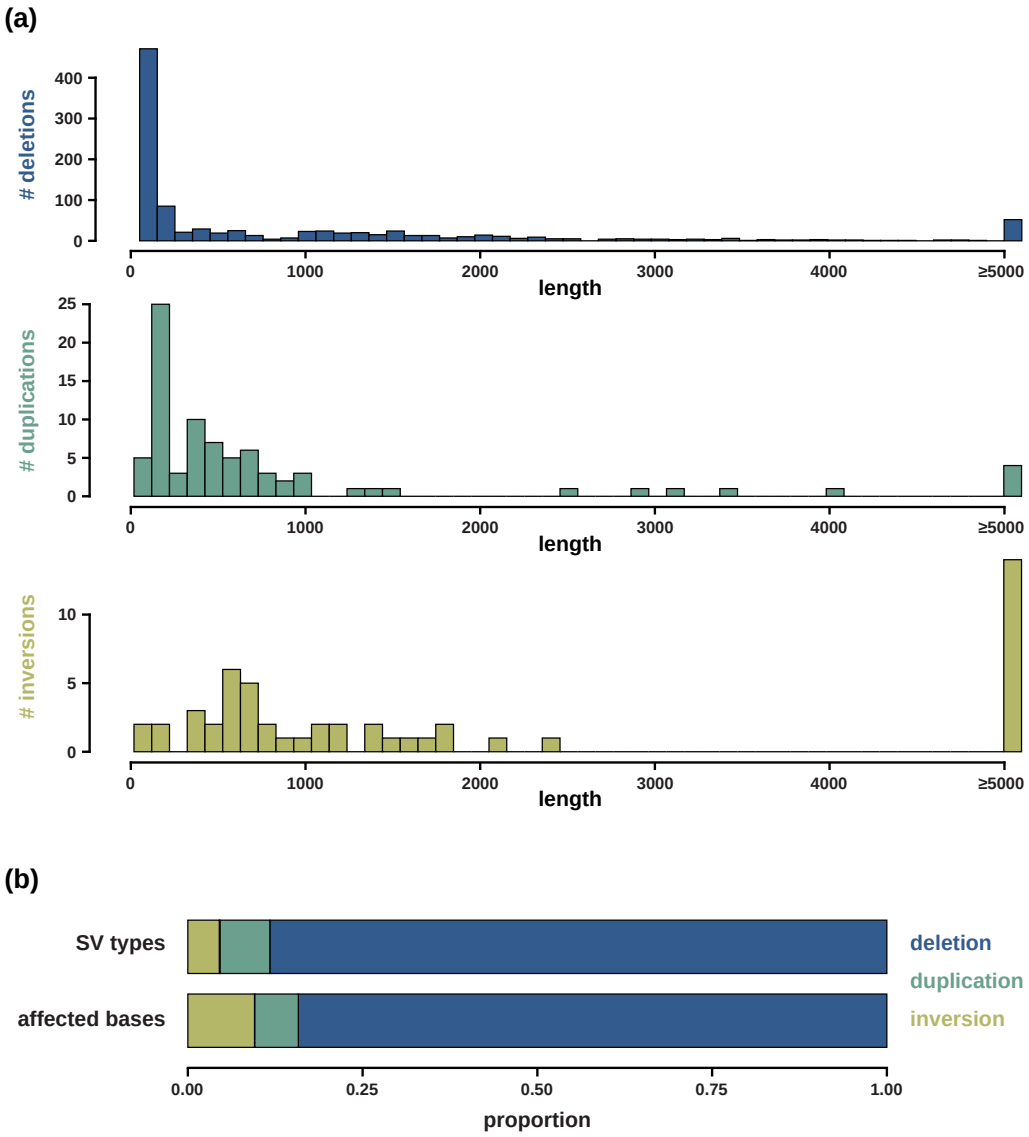

Figure 2. **Characteristics of structural variation in the aye-aye genome**. (a) Length distribution of structural variants (SVs; deletions are color-coded in blue, duplications in teal, and inversions in olive green; the single detected inversion is not shown). (b) Proportion of different SV types and base-pairs affected.

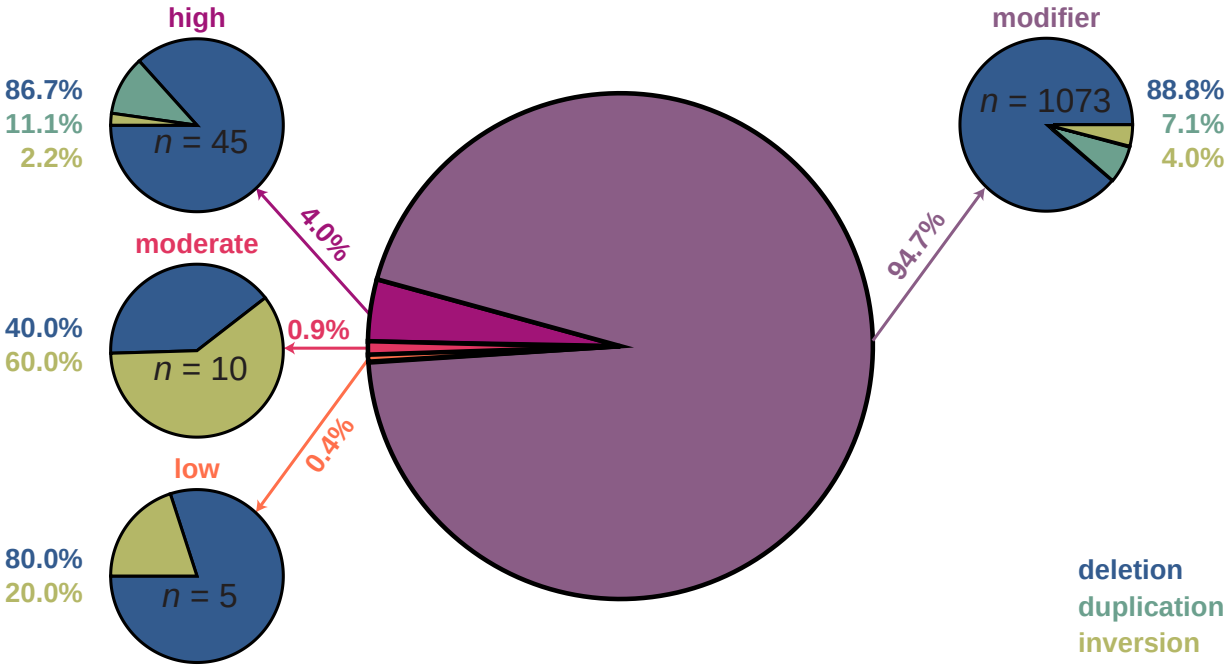

Figure 3. **Annotation of structural variation in the aye-aye genome**. The proportion of structural variants (deletions are color-coded in blue, duplications in teal, and inversions in olive green; the single detected inversion is not shown) classified as modifiers (shown in purple) as well as those predicted to have a high (pink), moderate (rose), and low (orange) impact.