Selection and local adaptation in capuchin monkeys revealed through fluorescence-activated cell sorting of feces (fecalFACS)

Joseph D. Orkin1,2*, Michael J. Montague3, Daniela Tejada-Martínez4,5, Marc de Manuel2, Javier del Campo6, Anthony Di Fiore7,8, Claudia Fontserè2, Jason A. Hodgson9,10, Mareike C. Janiak1,11, Lukas F.K. Kuderna2, Esther Lizano7,8, Yoshihito Niimura12, George H. Perry9,13, Jia Tang1, Wesley C. Warren14, Joao Pedro de Magalhães5, Shoji Kawamura15, Tomàs Marquès-Bonet2,16,17,18, Roman Krawetz19, Amanda D. Melin1,11,20*

1 Department of Anthropology and Archaeology, University of Calgary, Calgary, Canada
2 Institut de Biologia Evolutiva, Universitat Pompeu Fabra-CSIC, Barcelona, Spain
3 Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19146
4 Doctorado en Ciencias mención Ecología y Evolución, Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
5 Integrative Genomics of Ageing Group, Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool L7 8TX, UK
6 Department of Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL, USA
7 Department of Anthropology and Primate Molecular Ecology and Evolution Laboratory, University of Texas at Austin
8 College of Biological and Environmental Sciences, Universidad San Francisco de Quito, Cumbayá, Ecuador
9 Department of Anthropology, The Pennsylvania State University, State College, PA, USA
10 Department of Zoology, University of Cambridge, Cambridge, UK
11 Alberta Children’s Hospital Research Institute, Calgary, AB, Canada
12 Department of Applied Biological Chemistry, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan
13 Department of Biology, Huck Institute of Life Sciences, The Pennsylvania State University, State College, PA, USA
14 Division of Animal Sciences, School of Medicine, University of Missouri, Columbia, MO, 65211, USA
15 Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan
16 Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys, 23, 08010, Barcelona, Spain
17 CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain
18 Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, c/ Columns s/n, 08193 Cerdanyola del Vallès, Barcelona, Spain
19 Department of Cell Biology and Anatomy, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
20 Department of Cell Biology and Anatomy, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
Department of Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

* Corresponding authors

ABSTRACT

**Background**: Capuchins have the largest relative brain size of any monkey and a remarkable lifespan of 55 years, despite their small body size. Distributed widely across Central and South America, they are inventive and extractive foragers, known for their sensorimotor intelligence, dietary diversity, and ecological flexibility. Despite decades of research into their ecology and life history, little is known about the genomics of this radiation.

**Results**: We assemble a *de novo* reference genome for *Cebus imitator*, and provide the first genome annotation of a capuchin monkey. We identified 20,740 and 9,556 for protein- and non-coding genes, and recovered 23,402 orthologous groups. Through a comparative genomics approach across a diversity of mammals, we identified genes under positive selection associated with longevity and brain development, which are of particular relevance to capuchin and primate comparative biology. Additionally, we compared populations in distinct habitats, facilitated by our novel method for minimally-biased, whole-genome sequencing from fecal DNA using fluorescence activated cell sorting (FACS). By analyzing 23 capuchin genomes from tropical dry forest and rainforest, we identified population divergence in genes involved in water balance, kidney function, and metabolism, consistent with local adaptation to resource seasonality.

**Conclusions**: Our comparative study of capuchin genomics provides new insights into the molecular basis of brain evolution and longevity. These data also improve our understanding of processes of local adaptation to diverse and physiologically challenging environments. Additionally, we provide a technological advancement in use of non-invasive genomics to study free-ranging mammals through FACS.
BACKGROUND

Large brains, long lifespans, extended juvenescence, tool use, and problem solving are hallmark characteristics of the great apes, and are of enduring interest in studies of human evolution [1–6]. Similar suites of traits have arisen in other lineages, including some cetaceans, corvids and, independently, in another radiation of primates, the capuchin monkeys. Like great apes, they have diverse diets, consume and seek out high-energy resources, and engage in complex extractive foraging techniques [7,8] to consume difficult-to-access invertebrates and nuts [8]. Their propensity for tool use and their ecological flexibility may have contributed to their convergence with the great apes [9] offering opportunities for understanding their evolution via comparative methods [10–12].

Capuchins also offer excellent opportunities to study local adaptation to diverse habitats. While little is known about the migration, history, and population dynamics of this species, they presently range from Panama to northern Honduras [13–15], where they occupy a wide diversity of habitats, including rainforests and, in the northern extent of their range, tropical dry forests. Particular challenges of the tropical dry forest are staying hydrated during the seasonally prominent droughts, high temperatures in the absence of foliage, and coping metabolically with periods of fruit dearth (Figure 1). The sensory challenges of food search in dry versus humid biomes are also distinct. For example, odor detection and propagation is affected by temperature and humidity [16,17], and color vision is hypothesized to be adaptive in the search for ripe fruits and young reddish leaves against a background of thick, mature foliage [18], which
is absent for long stretches in dry deciduous forests. The behavioral plasticity of capuchins is widely acknowledged as a source of their ability to adapt to these dramatically different habitats [19–22]. However, physiological processes including water balance and metabolic adaptations to low caloric intake, and sensory adaptations to food search, are also anticipated to be targets of natural selection, as seen in other mammals [23–26].

Comparative genomics offers a unique opportunity to examine the molecular underpinnings of traits relevant to the evolution of humans and other great apes. Furthermore, the ecological flexibility of white-faced capuchins allows us to assess the influence of divergent and changing habitats on the process of adaptive evolution in these primates. In order to address the genetic underpinnings of capuchin adaptive evolution, we assembled the first reference genome of *Cebus imitator*. Additionally, to better understand the local adaptation of capuchins, we conducted high-coverage re-sequencing (10X - 47X) of 10 individuals from populations inhabiting distinct environments—one in a lowland evergreen rainforest (n=4), and another from a lowland tropical dry forest (n=6). Importantly, to facilitate the population-wide analyses without the need for potentially harmful invasive sampling of wild primates, we sequenced an additional 13 individuals at low coverage using a novel method for minimally-biased, whole-genome sequencing from fecal DNA using fluorescence-activated cell sorting (fecalFACS) that we developed (Figure 2). In our positive selection analyses, we focus on genes that may underlie sensation, cognition, and lifespan due to their relevance to capuchin-specific biology and adaptation. In our population comparison we predict genes related to water balance, metabolism will differ between dry forest and rainforest populations, reflecting local adaptation to different habitats. With respect to sensory systems, due to higher diversity of flora and fauna, including colorful fruits presented against green foliage, and increased humidity, genes underlying color vision (opsin genes) and chemosensation (olfactory and taste receptor genes) are predicted to be more diverse in rainforest habitats [27–29].
RESULTS

1. Comparative Genomics

Genome assembly and gene annotation

Our reference genome assembly for *Cebus imitator* is comprised of 7,742 scaffolds (including single contig scaffolds) with an N50 scaffold length of 5.2 Mb and an N50 contig length of 41 kb. The final ungapped assembly length is 2.6 Gb (GenBank accession: GCA_001604975.1). Our estimate of total interspersed repeats using WindowMasker [30] output is 45.8%. The numbers of annotated genes are 20,740 and 9,556 for protein-coding and non-coding genes, respectively (Table S1). Measures of gene representation using the known human RefSeq set of 56,230 transcripts show an average of >94% coverage with a mean identity of 92.5%. Overall, our draft assembly metrics and gene representation are consistent with other non-human primate (NHP) short-read reference assemblies [31].

Positive selection analyses

We recovered 23,402 Orthologous Groups (OGs). Capuchins share 18,475 OGs with human, 17,589 OGs with rhesus macaque, 15,582 OGs with mouse, and 14,404 OGs with dog. When we included orthologous genes that are present simultaneously in all 15 species, we recovered 7,519 OGs, which we subsequently used in the natural selection analyses (dN/dS=ω). We identified 612 genes under positive selection (p<0.05 after FDR correction) in the *Cebus* lineage using the branch model. We also performed a branch-site test using codeml in PAML[32] and identified a second set of 748 genes under positive selection in *Cebus* (Table S2 Sheet 2). The results of our enrichment analysis for biological processes using DAVID [33]
identified genes that underlie brain development, sensation, and lifespan, which were of particular interest given the derived features of capuchin biology (Figure 3).

**Brain development and longevity**

Capuchins have the largest brain to body ratio of any monkey, and are known for their sensorimotor intelligence [7] and derived cognitive abilities [8]. Of the 748 genes identified as being under positive selection in the branch-site model, 17 were previously associated with brain development (Table S2, Sheet 6, row 18) and 5 were linked to neurogenesis (Table S2, Sheet 6, row 116). For example, *WDR62, BPTF, BBS7, NUP113*, mutations are directly associated with brain size and related malformations, including microcephaly [34–37]. *MTOR* signaling malfunction is also implicated in developmental brain malformations [38], and *NUP113* is involved in nuclear migration during mammalian brain development [39]. Several genes are linked with brain tumor formation (including *ZNF217*), and others with cognitive ability (e.g. *PHF8* [40]).

We found 27 aging-related genes, as identified in the GenAge database [41,42], under positive selection in capuchins including *PARP1, MTOR, SREBF1, INSR, HTT, RB1* and *MDM2* (Table S2, Sheet 7). Of note, poly (ADP-ribose) polymerase 1 (PARP1) putatively serves as a determinant of mammalian aging due to its activity in the recovery of cells from DNA damage. In previous studies, gene expression levels of *PARP1* were inversely correlated with mammalian lifespan [43]. Another large body of research has associated the mechanistic target of rapamycin (*MTOR*) with aging and longevity in various organisms [44], making it a prime candidate for therapeutic interventions in aging [45]. *MTOR* acts as a regulator of cell growth and proliferation while also being generally involved in metabolism and other processes. Additional key genes in aging and metabolism include sterol regulatory element binding transcription factor 1 (*SREBF1*), which acts as a regulator of the metabolic benefits of caloric restriction [46,47], and the insulin
receptor (INSR), a major player in longevity regulation [48]. As for specific age-related diseases, huntingtin (HTT) is under selection in mammals; HTT is not only involved in Huntington’s disease but has also been associated with longevity in mice [49]. Lastly, various cell cycle regulators (e.g., RB1, MDM2) are also under positive selection in capuchins, and indeed, cell cycle is an enriched term among positively selected genes (Table S2), though these could be related to other life history traits like developmental schedules that correlate with longevity.

2. Population Genomics and Local Adaptation

Population structure, genetic diversity, and demographic history in Costa Rican white-faced capuchins

Of the 24 capuchin DNA samples that we sequenced and mapped to our reference genome, 15 were fecal-derived (Table S3). When comparing the high coverage tissue-derived genomes from the Santa Rosa site to those generated from our novel application of fluorescence-activated cell sorting to isolate fecal-sourced cells (fecalFACS), we observed no substantial difference in quality, coverage, heterozygosity, or GC content (Figures 2, S1, S2, Supplemental Text). This includes the first (to our knowledge) high-coverage (12.2 X) whole mammalian genome generated from a fecal sample.

The pattern of clustering in our maximum likelihood single nucleotide variant (SNV) tree recapitulates the expected patterns of geographic distance and ecological separation in our samples (Figure 4). Likewise, in the projected PCA all individuals from the seasonal dry forests in the northwest are sharply discriminated from individuals inhabiting the southern rainforests along PC1. These relationships are not perturbed by depth of coverage, or source material (tissue-based vs fecalFACS genomic libraries) (Figure 4). Levels of heterozygosity calculated in overlapping 1 Mb genomic windows (with a step size of 100 kb) were significantly higher in the
southern population (W = 1,535,400,000, p-value < 2.2e-16; Figure 5A, Figure S3).

Furthermore, the median pairwise heterozygosity for each southern individual (range: 0.00065 - 0.00071) was higher than any northern monkey (0.00047 - 0.00057) (W = 0, p-value = 0.009524; Table S5). In the northern population, we also identified long runs of homozygosity significantly more often (W = 24, p-value = 0.009524), and more of the longest runs (>= 5 Mb) (W = 1315.5, p-value = 0.03053; Figures 5B, S4, S5). Pairwise sequential Markovian coalescent (PSMC) analysis of demographic history (Figure 5C) reveals that white-faced capuchins had a peak effective population size of ~60,000 effective individuals ~1 mya, which declined to fewer than 20,000 during the middle to late Pleistocene transition. After recovering during the middle Pleistocene, they declined precipitously through the late Pleistocene to fewer than 5,000 effective individuals.

Local adaptation to seasonal dry forest biome

We predicted that genes related to water balance, metabolism, and sensation would differ between dry forest and rainforest populations, reflecting local adaptation to different habitats. To test this, we searched for associations between genes in windows with high F_{ST} and divergent non-synonymous SNVs with high or moderate effect. Of the 299 genes identified in high F_{ST} windows, 39 had highly differentiated non-synonymous SNVs (Figure S6, Tables S6, S7). Our enrichment analysis identified a single significant GO biological process: regulation of protein localization to cilium (GO: 1903564). Remarkably—and in accordance with our hypothesis—disruptions of cilia proteins are predominantly associated with disorders of the kidney and retina (ciliopathies) [50]. Furthermore, enrichment of genes associated with disease states was linked to kidney function, metabolism, and muscular wasting. These genes are good candidates for adaptive resilience to seasonal water and food shortages and warrant further investigation. We highlight several genes of particular promise (Figure 6, Figure S7).
Evidence of adaptation to food and water scarcity

Multiple candidate genes indicate that dry-forest capuchins could be adapted to seasonal drought-like conditions and food scarcity. SERPINC1 encodes antithrombin III, which is involved in anticoagulant and anti-inflammatory responses associated with numerous kidney-related disorders including salt-sensitive hypertension, proteinuria, and nephrotic syndrome [51–56]. SERPINC1 is overexpressed in the renal cortex of Dahl salt-sensitive rats fed a high salt diet [55], suggesting that variants could be facilitating water/salinity balance in the dry-forest capuchins. Additionally, sequence variants of AXDND1 (as identified in the GeneCards database) are associated with nephrotic syndrome, a kidney disorder that commonly presents with edema and proteinuria. BCAS3 is expressed in multiple distal nephron cells types [57], and is associated with four pleiotropic kidney functions (concentrations of serum creatinine, blood urea nitrogen, uric acid, and the estimated glomerular filtration rate based on serum creatinine level) [58]. Although we did not identify any population-specific non-synonymous SNVs in BCAS3, the genomic windows encompassing the gene rank among the highest regions of FST in our dataset, and intronic variation in BCAS3 putatively impacts estimated glomerular filtration rate in humans [57].

The association of BCAS3 and serum creatinine is of particular interest, given field observations of capuchins from SSR, whose urinary creatinine levels are associated with decreased muscle mass during periods of seasonally low fruit availability [59]. Creatinine is a byproduct of the metabolism of creatine phosphate in skeletal muscle, which is normally filtered by the kidneys, and can be used as a clinical biomarker of kidney function, chronic kidney disease [58], and as a monitor of muscle mass [59,60]. Multiple congenital muscular dystrophies—including mild forms that present with muscular wasting—and abnormal circulating creatine kinase concentration (HP:0040081) are associated with candidate genes ITGA7, ISPD (CRPPA), and SYNE2 [61–64]. Furthermore, transgenic overexpression of ITGA7 has been
shown to reduce muscular pathologies caused by mutations in LAMA2 [65], which falls in a high F\textsubscript{ST} window. The relationship among seasonal resource availability, kidney function, and muscle mass is further supported by a potential adaptive role in capuchin sugar metabolism and frugivory. In particular, GLIS3 is one of two genes known to be associated with type 1 diabetes, type 2 diabetes, and neonatal diabetes [66], and appears to be diverging between populations. Additionally, while insulin receptor substrate (IRS4) did not fall in a high F\textsubscript{ST} window, we observed a non-synonymous SNP of medium effect fixed between populations. Given the appearance of both diabetes and kidney disorders in our gene sets, we conducted an \textit{a posteriori} search of our high F\textsubscript{ST} gene set for overrepresentation of genes associated with diabetic nephropathy (EFO\textunderscore 0000401) in the GWAS catalogue (Table S6). Seven genes were present in both our gene set of 299 high F\textsubscript{ST} genes and the diabetic nephropathy set of 117. Given the 16,553 annotated genes with HGNC Gene IDs, seven overlapping genes would occur with p = 0.00046 when permuted 100,000 times. We take this as promising evidence that these genes have been under selection in the northwestern population.

\textbf{Evidence of adaptation in sensory systems}

Given the ecological differences between the northern dry and southern wet forests, we predicted that the evergreen, humid environment of the lowland rainforest would favor enhanced diversity of both color vision (opsin genes) and olfaction (olfactory receptor genes) would result in population specific variation in chemosensory and visual genes. We did not find support for this; in both populations, we observed similar polymorphism at each of the three medium/long wave cone opsin tuning sites (180, A/S; 277, F/Y; and 285, T/A) (Table S8). None of these codons is a novel variant [67,68], providing no support for the hypothesis of differences in the perception of photopic (cone-driven) vision between the two biomes. However, we did observe some evidence for population specific variation associated with the photoreceptive layers of the
retina. First, creatine kinase plays an important role in providing energy to the retinal pigment epithelium [69]. Secondly, we identified a fixed non-synonymous SNV in CCDC66, which falls in a high F\textsubscript{ST} region and is heavily expressed in photoreceptive layers of the retina. Electroretinography of CCDC66 -/- mice reveals a significant reduction in scotopic (rod-driven) photoreceptor response [70], indicating a potential effect on vision in low-light conditions. Curiously, CCDC66 -/- mice, also display neurodegeneration of the olfactory bulb, and have reduced odor discrimination performance of lemon smells [71]. Turning to olfaction, we identified 614 olfactory receptor (OR) genes and pseudogenes in the capuchin reference genome (408 intact, 45 truncated, and 161 pseudogenized (Table S9). To test for population differences in the OR gene repertoire, we assembled each olfactory gene/pseudogene independently in each individual. The proportion of total functional ORs was stable across individuals and populations, with trivial fluctuations (North $\bar{x}$ = 411, s = 1.6; South $\bar{x}$ = 408.5, s = 1.3), possibly driven by a small difference in OR family 5/8/9 (Table S10). We also identified 8 vomeronasal and 28 taste receptor and taste receptor-like genes (Table S11), two of which (TAS1R and TAS2R4) have non-synonymous SNVs with fixed variants in the north (Table S12). The functional significance of these variants is unknown, but may be revealed via cellular expression systems in future research [72].

**DISCUSSION**

**Comparative genomics of white-faced capuchins**

Among primates, capuchin monkeys are known for their relatively large brains, cognitive capacity, and sensorimotor intelligence [7,8,73,74]. Accordingly, it is perhaps unsurprising to see positive selection in the *Cebus* lineage and evidence of shifts in gene function linked to brain function and development relative to other primates. In particular, positive selection in
WDR62, BPTF, BBS7, NUP133, and MTOR, and PHF8 supports our hypothesis that the capuchin lineage has undergone adaptation linked to brain development. The association of several of these genes with size related brain malformations [34–37], such as microcephaly, suggests that they could be influencing the large relative brain size of capuchins. Furthermore, the evidence of selection in PHF8, which is associated with human cognitive capacity, aligns with the link between brain size and intelligence that has been observed in other primates [75]. While we highlight here the putative functional roles of these genes, which are based on clinical studies and comparative genomics, we acknowledge that further examination of their function in the context of capuchin biology is warranted.

In the context of longevity, it is noteworthy that we observed genes under selection associated with DNA damage response, metabolism, cell cycle and insulin signaling [76]. Of particular interest are: PARP1, MTOR, SREBF1, INSR1, and HTT. Damage to the DNA is thought to be a major contributor to aging [77]. Previous studies have also shown that genes involved in DNA damage responses exhibit longevity-specific selection patterns in mammals [41]. It is therefore intriguing that PARP1, a gene suggested to be a determinant of mammalian aging [43], is under selection in capuchins. Other genomes of long-lived mammals also revealed genes related to DNA repair and DNA damage responses under selection [78,79]. In the context of longevity, it is also noteworthy that we observed genes under selection associated with metabolism, cell cycle and insulin signaling. Other genome sequencing efforts of long-lived mammals also revealed changes in such pathways [79,80]. Intriguingly, short-lived species also exhibit genes under selection related to insulin receptors, raising the possibility that the same pathways associated with aging in model organisms are involved in the evolution of both short- and long lifespans [81], an idea supported by our results. Of course, because aging-related genes often play multiple roles, for example in growth and development, it is impossible to tell for sure whether selection in these genes is related to aging or to other life-history traits, like growth rates and developmental times, that in turn correlate with longevity [82]. Therefore,
although we should be cautious about the biological significance of our findings, it is tempting to speculate that, like in other species, changes to specific aging-related genes or pathways, could contribute to the longevity of capuchins.

**Population genomics and local adaptation with fecalFACS**

Through a novel use of flow cytometry/FACS, we have developed a new method for the isolation of epithelial cells from mammalian feces for population genomics. We generated the first high-coverage, minimally biased mammalian genome solely from feces. Additionally, we have demonstrated that fecalFACS can be used to generate low coverage SNP datasets that are suitable for population assignment and clustering. FecalFACS is cost-effective and minimizes the biases that commonly occur in traditional bait-and-capture approaches to the enrichment of endogenous DNA from feces. Furthermore, fecalFACS does not require costly impractical preservation of biomaterial in liquid nitrogen; rather, we rely on room-temperature stable storage in RNAlater. FecalFACS offers great benefits to the field of mammalian conservation and population genomics.

White-faced capuchins are the most northerly distributed member of the Cebinae, having dispersed over the isthmus of Panama in a presumed speciation event with *C. capucinus* in South America ~1.6 mya [13–15]. After expanding during the early late Pleistocene, white-faced capuchins appear to have undergone a dramatic reduction in effective population size. This pattern predates the movement of humans into Central America, and could reflect a series of population collapses and expansions caused by glacial shifts and fluctuating forest cover availability during the Pleistocene. At a finer scale, we observed a clear demarcation between the northern dry- and southern wet-forest populations in Costa Rica. Higher levels of heterozygosity in the south and lower levels in the northwest are in accordance with the hypothesis that capuchins dispersed northwards across Costa Rica. White-faced capuchins in
SSR are near the northernmost limits of their range, which extends as far north as Honduras, and we predict they may represent some of the least genetically diverse members of their species. Given the limitations of the available sampling sites, it is possible that the appearance of an ecological divide is actually evidence of isolation by distance; however, given that the single individual from Cañas clusters closely with the individuals from SSR, despite a geographic distance of more than 100 km, we suggest that isolation by distance does not completely explain the population differentiation.

We found evidence that the recent northern expansion of *Cebus imitator* has undergone local adaptation to the extreme seasonality of rainfall and food availability. The effects of seasonality have been linked to biological consequences for Costa Rican white-faced capuchins in other contexts. Seasonal fluctuations in food abundance and rainfall are associated with compositional changes in the gut microbiome of capuchins at SSR [83], which differs markedly from that observed in capuchins inhabiting nonseasonal forests [84]. Previous capuchin research in the dry forest also demonstrates seasonal negative energy balance and periods of pronounced muscle loss through catabolic processes [59]. These observations fit with the notion that animals living in seasonal environments, or pursuing seasonal migrations, are more likely to have weight fluctuations through binge-subsist cycles that map onto food abundance [85]. Accordingly, we observed population-specific variation in genes implicated in water/salt balance, kidney function, muscle wasting, and metabolism (e.g. *SERPINC1*, *BCAS3*, *ITGA7*, *ISPD*, and *GLIS3*). In light of this, we contend that seasonal drought and food shortages would create an environment favoring efficient catabolism when needed and adaptations for maintaining water balance. Given that selection operates on both gene function and regulation, we suspect the observed variation is affecting gene expression or enzymatic efficiency, which offers a promising avenue for future research.

Additionally, we found evidence of 408 OR genes, 28 TASR, and 7 VR genes that are putatively functional. These numbers are similar to, or slightly higher than, the number of
chemosensory genes identified in other anthropoid primates [86–90]. The VR gene repertoire of capuchins highlights the persistent role of the vomeronasal organ that is used in social communication of other mammals, but that has been nearly lost from all African and Asian monkeys [89]. Like most other primates in the Americas, capuchins possess an intriguing color vision system characterized by extensive intraspecific genotypic and phenotypic variation [32–35]. Contrary to our prediction, dry-forest and rainforest capuchins have similar numbers of color vision, taste and olfactory receptors. This indicates that there are not greater numbers of functional sensory genes in areas of higher vegetative biodiversity or humidity. The tuning of the chemosensory receptors may vary between habitat types and may be elucidated by future work. While we did observe population-specific variation in CCDC66, which is expressed in photoreceptive layers of the retina and may affect odor discrimination, the ecological significance of this result, if any, is unclear at present but may warrant future attention.

CONCLUSION

We provide the first annotated reference assembly for a capuchin monkey. We observed evidence of selection in *C. imitator* on genes involved with brain function and cognitive capacity. These results are in accordance with the remarkably long life span, large brain, and high degree of sensorimotor intelligence that has been observed in capuchins. These genes are good candidates for further investigation of traits which have evolved in parallel in apes and other mammals. Through a novel use of flow cytometry/FACS, we developed a new method for the isolation of epithelial cells from mammalian feces for population genomics. FecalFACS allowed us to generate both the first high-coverage, minimally biased mammalian genome solely from feces, as well as low coverage SNV datasets for population level analyses. In our population level analysis of wet- and dry-forest capuchins, we observed both evidence of population structure between and local adaptation to these different habitats. In particular, we identified
selection in genes related to food and water scarcity, as well as muscular wasting, all of which have been observed during seasonal extremes in the dry forest population.

METHODS

Study populations and sample collection

Central American white-faced capuchins (*Cebus imitator*), a member of the gracile radiation of capuchins (genus *Cebus*) [91], were recently recognized as a species, distinct from *C. capucinus* in South America [14]. *Cebus imitator* occupies a wide diversity of habitats, spanning lowland rainforests and cloud forests in Panama and throughout southern, eastern, and central Costa Rica, and tropical dry forests in northwestern Costa Rica and Nicaragua. The annual precipitation and elevation of rainforest versus dry forest biomes in their current range vary dramatically, leading to considerable variation in the resident flora and fauna [92,93]. We sampled individual Costa Rican capuchin monkeys from populations inhabiting two distinct habitats. 1) lowland rainforest around Quepos, Puntarenas Province; and 2) tropical dry forest at two sites in Guanacaste Province. In total, we collected samples from 23 capuchins, a list of which is provided in table S3.

We sampled capuchins inhabiting a lowland tropical rainforest biome by collaborating with *Kids Saving the Rainforest* (KSTR) in Quepos, Costa Rica. We acquired blood samples from 4 wild capuchins from nearby populations who were undergoing treatment at the facility (although we were unable to collect paired fecal samples). For one of these individuals, an adult male white-faced capuchin that was mortally wounded by a vehicle in Costa Rica, we additionally sampled tissues from several organs. DNA derived from the kidney was used for the reference genome assembly.

We collected 21 samples from 19 individuals in the northern tropical dry forest. 16 fecal samples and 4 tissue samples were from free-ranging white-faced capuchin monkeys (*Cebus*
imitator) at in the Sector Santa Rosa (SSR), part of the Área de Conservación Guanacaste in northwestern Costa Rica, which is a 163,000 hectare tropical dry forest nature reserve (Figure 1). Behavioral research of free-ranging white-faced capuchins has been ongoing at SSR since the 1980’s which allows for the reliable identification of known individuals from facial features and bodily scars [94]. The 16 fresh fecal samples were collected from 14 white-faced capuchin monkeys immediately following defecation (Table S3). We placed 1 mL of feces into conical 15 mL tubes pre-filled with 5 mL of RNAlater. RNAlater preserved fecal samples were sent to the University of Calgary, where they were stored at room temperature for up to three years. To evaluate other preservation methods, we also collected two additional capuchin monkey fecal samples (SSR-FL and a section of SSR-ML), which we stored in 1X PBS buffer and then frozen in liquid nitrogen with a betaine cryopreservative [95]. Given the logistical challenges of carrying liquid nitrogen to remote field sites, we prioritized evaluation of samples stored in RNAlater. We also collected tissue and blood samples opportunistically. During the course of our study, 4 individual capuchin monkeys died of natural causes at SSR, from whom we were able to collect tissue samples, which were stored in RNAlater. Additionally, we collected a blood sample from 1 northern dry-forest individual housed at KSTR that originated near the town of Cañas, which is ~100 km southwest of SSR.

**Genome-wide sequencing, genome assembly and gene annotation**

We assembled a reference genome for *Cebus imitator* from DNA extracted from the kidney of a male Costa Rican individual (KSTR64) using a short read approach (Illumina HiSeq 2500). Based on a genome estimate of 3 Gb, the total sequencing depth generated was 81X, including 50X of overlapping read-pairs (200 bp insert), 26X and 5X of 3 and 8 kbs insert read pairs, respectively. The combined sequence reads were filtered and assembled using default parameter settings with ALLPATHS-LG [96]. To improve the quality of gene annotation, we isolated total RNA from the whole blood of an adult male white-faced capuchin (ID: CNS-HE).
permanently residing at the KSTR wildlife rehabilitation center. The blood was immediately
stored in a PAXgene blood RNA tube (Qiagen), and frozen at ultralow temperatures for
subsequent use. To extract total RNA, we used the PAXgene Blood RNA kit following the
manufacturer’s recommended protocols. A RiboZero library construction protocol was followed
according to the manufacturer’s specifications and sequenced on an Illumina HiSeq 2000
instrument creating 150 bp paired-end reads. We assembled the FASTQ sequence files into
transcripts with Trinity, and submitted the assembled transcriptome to the National Center for
Biotechnology Information (NCBI) to assist in gene annotation. The capuchin genome assembly
was annotated with the NCBI pipeline previously described here:
(http://www.ncbi.nlm.nih.gov/books/NBK169439/).

**Phylogenetic arrangement and data treatment**

The phylogenetic arrangement in this study included 14 species as outgroups to C.
imitator: three Platyrhini (*Callithrix jacchus, Aotus nancymaeae, Saimiri boliviensis*), six
Catarrhini (*Macaca mulatta, Rhinopithecus roxellana, Nomascus leucogenys, Pan troglodytes,*
*Homo sapiens, Pongo abelii*), one Strepsirrhini (*Microcebus murinus*), one rodent (*Mus
*musculus*), and three Laurasiatheria (*Canis lupus familiaris, Bos taurus, and Sus scrofa*).
Genomic cds were downloaded from Ensembl and NCBI (Table S13). The sequences per
genome were clustered using CD-HITest version 4.6 [97] with a sequence identity threshold of
90% and an alignment coverage control of 80%. To remove low quality sequences and keep the
longest transcript per gene, we used TransDecoder.LongOrfs and TransDecoder.Predict
(https://transdecoder.github.io) with default criteria.

**Orthology identification**

The orthology assessment was performed with OMA stand-alone v. 2.3.1 [98]. The OMA
algorithm makes strict pairwise “all-against-all” sequence comparisons and identifies the
orthologous pairs (genes related by speciation events) based on evolutionary distances. These orthologous genes were clustered into Orthologous Groups (OGs). All OGs included one ortholog sequence from capuchin and at least one outgroup. The tree topology was obtained from TimeTree (http://www.timetree.org/). We identified 7,519 OGs present among the 15 species. Each orthogroup shared by all species was translated into amino acids using the function pxtlate -s in phyx [99]. Amino acid sequences were aligned using the L-INS-i algorithm from MAFFT v.7 [100]. We generated codon alignments using pxaa2cdn in phyx. To avoid false positives in low quality regions, the codon alignments were cleaned with the codon.clean.msa algorithm in rphast [101], using human as a reference sequence. We used conservative methodologies of homology and data cleaning to obtain a smaller number of orthologous genes that avoided false positives with high confidence.

Positive natural selection analysis through codon-based models of evolution and enrichment tests

To evaluate signals consistent with positive selection in the C. imitator genome, we explored variation in the ratio of non-synonymous and synonymous substitutions ($d_N/d_S=\omega$) in the ancestor of Cebus. We used branch and branch-site substitution models with a maximum likelihood approach in PAML v4.9 [32], which we implemented through the python framework ETE-toolkit with the ete-evol function [102]. We compared the null model where the omega ($\omega$) value in the branch marked as foreground was set with 1, with the model where the $\omega$ value was estimated from the data [103]. Likelihood ratio tests (LRT) were used to test for significance between the models and probability values were adjusted with a false discovery rate correction for multiple testing with a q-value < 0.05 for the two positive selection models (branch and branch-site).

We performed functional annotation analysis using DAVID bioinformatics resources - DAVID 6.7 [33], to ascertain which ontology processes the genes with signals of positive
selection were involved. We focussed the enrichment analysis on two functional categories: Biological Processes (BP) and Genetic Association Database (GAD). GAD is a database from complex diseases and disorders in humans. Finally, the genes with positive section signal were intersected with the GenAge database (build 19,307 genes) [42].

**FACS**

Before isolating cells by fluorescence-activated cell sorting, fecal samples were prepared using a series of washes and filtration steps. Fecal samples were vortexed for 30 s and centrifuged for 30 s at 2,500 g. Then the supernatant was passed through a 70 μm filter into a 50 mL tube and washed with DPBS. After transferring the resultant filtrate to a 15 mL tube, it was centrifuged at 1,500 RPM for 5 minutes to pellet the cells. Then we twice washed the cells with 13 mL of DPBS. We added 500 μL of DPBS to the pellet and re-filtered through a 35 μm filter into a 5 mL FACS tube. We prepared a negative control (to control for auto-fluorescence) with 500 μL of DPBS and one drop of the cell solution. To the remaining solution, we added 1 μL of AE1/AE3 Anti-Pan Cytokeratin Alexa Fluor® 488 antibody or TOTO-3 DNA stain, which we allowed to incubate at 4°C for at least 30 minutes.

We isolated cells using a BD FACSARia™ Fusion (BD Biosciences) flow cytometer at the University of Calgary Flow Cytometry Core. To sterilize the cytometer's fluidics before processing each sample, we ran a 3% bleach solution through the system for four minutes at maximum pressure. We assessed background fluorescence and cellular integrity by processing the negative control sample prior to all prepared fecal samples. For each sample we first gated our target population by forward and side scatter characteristics that were likely to minimize bacteria and cellular debris (Figure S8). Secondary and tertiary gates were implemented to remove cellular agglomerations. Finally, we selected cells with antibody or DNA fluorescence greater than background levels. In cases when staining was not effective, we sorted solely on the first three gates. Cells were pelleted and frozen at -20°C.
DNA Extraction and Shotgun Sequencing

We extracted fecal DNA (fDNA) with the QIAGEN DNA Micro kit, following the "small volumes of blood" protocol. To improve DNA yield, we increased the lysis time to three hours, and incubated 50 µL of 56°C elution buffer on the spin column membrane for 10 minutes. DNA concentration was measured with a Qubit fluorometer. Additionally, to calculate endogenous DNA enrichment, we extracted DNA directly from five fecal samples prior to their having undergone FACS. We extracted DNA from the nine tissue and blood samples using the QIAGEN Gentra Puregene Tissue kit and DNeasy blood and tissue kit, respectively.

For the fecal samples, DNA was fragmented to 350 bp with a Covaris sonicator. We built whole genome sequencing libraries with the NEB Next Ultra 2 kit using 10-11 PCR cycles. Fecal genomic libraries were sequenced on an Illumina NextSeq (2x150 PE) at the University of Calgary genome sequencing core and an Illumina HighSeq 4000 at the McDonnell Genome Institute at Washington University in St. Louis (MGI). Using ½ of one HiSeq 4000 lane, we achieved an average coverage of 12.2X across the Cebus imitator 1.0 genome (sample SSR-ML). Other fecal samples were sequenced to average depths of 0.1-4.4X (Table S4). High-coverage (10.3-47.6X), whole genome shotgun libraries were prepared for the blood and tissue DNA samples and sequenced on an Illumina X Ten system at MGI. For population analyses within capuchins, we mapped genomic data from all 23 individuals sequenced (Table S3) to the reference genome.

Mapping and SNV Generation

Reads were trimmed of sequencing adaptors with Trimmomatic [104]. Subsequently, we mapped the Cebus reads to the Cebus imitator 1.0 reference genome (GCF_001604975.1) with BWA mem [105] and removed duplicates with Picard Tools (http://broadinstitute.github.io/picard/) and SAMtools [106]. We called SNVs for each sample.
independently using the *Cebus* genome and the GATK UnifiedGenotyper pipeline (*-out_mode EMIT_ALL_SITES*) [107]. Genomic VCFs were then combined using GATK’s CombineVariants restricting to positions with a depth of coverage between 3 and 100, mapping quality above 30, no reads with mapping quality zero, and variant PHRED scores above 30. Sequencing reads from one of the high coverage fecal samples (SSR-FL) bore a strong signature of human contamination (16%), and were thus excluded from SNV generation. We included reads from nine tissue/blood samples and one frozen fecal sample with high coverage (SSR-ML). In total, we identified 4,184,363 SNVs for downstream analyses.

To remove potential human contamination from sequenced libraries, we mapped trimmed reads to the *Cebus imitator* 1.0 and human (hg38) genomes simultaneously with BBsplit [108]. Using default BBsplit parameters, we binned separately reads that mapped unambiguously to either genome. Ambiguously mapping reads (i.e. those mapping equally well to both genomes) were assigned to both genomic bins, and unmapped reads were assigned to a third bin. We calculated the amount of human genomic contamination as the percentage of total reads unambiguously mapping to the human genome (Table S4). After removing contaminant reads, all libraries with at least 0.5X genomic coverage were used for population structure analysis.

In order to test the effect of fecalFACS on mapping rates, we selected five samples at random (SSR-CH, SSR-NM, SSR-LE, SSR-PR, SSR-SN) to compare pre- and post-FACS mapping rates. To test for an increase in mapping percentage, we ran a one-sample paired Wilcoxon signed-rank test on the percentages of reads that mapped exclusively to the *Cebus* genome before and after FACS. Additionally, we ran Pearson’s product moment correlations to test for an effect of the number of cells (log10 transformed) on rates of mapping, read duplication, and nanograms of input DNA. The above tests were all performed in R.

**High coverage fecal genome comparison**
We made several comparisons between our high-coverage feces-derived genome and the blood/tissue-derived genomes using window-based approaches. For each test, the feces-derived genome should fall within the range of variation for members of its population of origin (SSR). Deviations from this, for example all fecal genomes clustering together, would indicate biases in our DNA isolation methods. To assess this, we constructed 10 KB windows with a 4KB slide along the largest scaffold (21,314,911 bp) in the C. imitator reference genome. From these windows, we constructed plots of coverage density and the distribution of window coverage along the scaffold. Secondly, we assessed the level of heterozygosity in 1 MB / 200 KB sliding windows throughout the ten largest scaffolds. For each high-coverage genome, we plotted the density distribution of window heterozygosity. We measured genome-wide GC content with the Picard Tools CollectGcBiasMetrics function. The percentage of GC content was assessed against the distribution of normalized coverage and the number of reads in 100 bp windows per the number reads aligned to the windows.

**Population Structure**

Given the large degree of difference in coverage among our samples, (less than 1X to greater than 50X), we performed pseudodiploid allele calling on all samples. For each library, at each position in the SNV set, we selected a single, random read from the sequenced library. From that read, we called the variant information at the respective SNV site for the given library. In so doing, we generated a VCF with a representative degree of variation and error for all samples.

To assess population structure and infer splits between northern and southern groups of Costa Rican white-faced capuchins, we constructed principal components plots with EIGENSTRAT [109] and built population trees with TreeMix [110]. Because we ascertained variants predominantly with libraries that were of tissue/blood origin, we built principal components solely with SNVs from these libraries and projected the remaining fecal libraries
onto the principal components. For our maximum likelihood trees, we used two outgroups
(Saimiri sciureus, and Cebus albifrons), with S. sciureus serving as the root of the tree. Given
the geographic distance and anthropogenic deforestation between northern and southern
populations, we assumed no migration. To account for linkage disequilibrium, we grouped SNVs
into windows of 1,000 SNVs. Population size history was inferred from the highest coverage
non-reference individual, SSR-RM08 with PSMC [111], using default parameters

Local adaptation, $F_{ST}$, heterozygosity, relatedness

For all analyses of local adaptation and heterozygosity between populations, we
excluded individuals from our low coverage dataset. We tested for the degree of relatedness
among all high and low coverage individuals using READ [112] and identified two of the high-
coverage individuals from SSR, SSR-ML and SSR-CR, as potential first degree relatives (Figure
S9, Table S14). For all statistical analyses of high-coverage samples, we removed SSR-ML,
because SSR-CR was sequenced to higher average depth.

For each individual, we calculated heterozygosity in 1 Mb / 200 Kb sliding windows
across the genome for all scaffolds at least 1 Mb in length. Windows were generated with
BedTools windowMaker [113] and heterozygosity was calculated as the per-site average within
each window. Based upon a visual inspection of the average heterozygosity values across the
genome (Figure 5), we classified a window as part of a run of homozygosity if the window's
average heterozygosity fell below 0.0002. Descriptive statistics and two-sided Wilcoxon tests
were calculated in R.

For each high-coverage sample, we calculated the Hudson's $F_{ST}$ ratio of averages [114]
in 20 kb windows with a slide of 4 Kb across the genome. Among the genes present in each
window in the top 0.5% and top 0.1% of $F_{ST}$ values, we searched for SNPs with high or
moderate effects using SnpEff and identified those SNPs with high $F_{ST}$ values (> 0.75) using
VCFtools. We searched for functional enrichment of our population gene set using ToppFun in the ToppGene Suite [115]. In an effort to identify candidate genes for further investigation, we set low thresholds for significance (FDR p-value < 0.1 and the minimum number of genes per category to 1).

**Chemosensory genes**

The chemosensory behaviors of capuchins have been well-studied [31], and taste and olfaction are suspected to play an important role in their foraging ecology. **614 orthologous olfactory** receptor genes were identified in the *Cebus* reference genome using previously described methods [116]. Briefly, putative OR sequences were identified by conducting TBLASTN searches against the capuchin reference assembly using functional human OR protein sequences as queries with an e-value threshold of 1e-20. For each reference-derived OR gene, we added 500 bp of flanking sequence to both the 5′ and 3′ ends in a bed file. For each individual, we extracted the OR gene region from the gVCF and generated a consensus sequence defaulting to the reference allele at variable site using bcftools [117]. The number of intact, truncated, and pseudogenized OR genes in each individual were identified using the ORA pipeline [118]. We considered an OR gene to be putatively functional if its amino acid sequence was at least 300 amino acids in length. ORA was further used to classify each OR into the appropriate class and subfamily with an e-value cutoff of 1e-10 to identify the functional OR subgenome for each individual [118,119]. Taste and vomeronasal receptor genes were identified in the NCBI genome annotation, and variable positions were located by scanning the VCF with VCFtools [120] and bash. The positions of *Cebus* opsin tuning sites have been identified previously [68]. With the high-coverage dataset, we identified the allele(s) present at each locus in the VCF. For each low-coverage fecal-derived genome, we located the position of the tuning site in the bam file using SAMtools [117] tview and manually called the variant when possible.
DECLARATIONS

Ethics approval and consent to participate

This research adhered to the laws of Costa Rica, the United States, and Canada and complied with protocols approved by the Área de Conservación Guanacaste and by the Canada Research Council for Animal Care through the University of Calgary’s Life and Environmental Care Committee (ACC protocol AC15-0161). Samples were collected with permission from the Área de Conservación Guanacaste (ACG-PI-033-2016) and CONAGEBIO (R-025-2014-OT-CONAGEBIO; R-002-2020-OT-CONAGEBIO). Samples were exported from Costa Rica under permits from CITES and Area de Conservacion Guanacaste (2016-CR2392/SJ #S 2477, 2016-CR2393/SJ #S 2477, DGVS-030-2016-ACG-PI-002-2016; 012706) and imported with permission from the Canadian Food and Inspection Agency (A-2016-03992-4).

Consent for Publication

Not applicable

Availability of data and materials

The reference genome is available at NCBI through BioProjects PRJNA298580 and PRJNA328123. RNAseq reads used in genome annotation can be accessed through PRJNA319062. The sequencing reads used in the local adaptation will be released by NCBI upon publication (PRJNA610850), and are available immediately to reviewers and editors upon request from the corresponding author.
**Competing Interests**

The authors declare that they have no competing interests.

**Funding**

Funding was provided by Washington University in St. Louis, the Canada Research Chairs Program, and a National Sciences and Engineering Council of Canada Discovery grant (to A.D.M.), the Alberta Children’s Hospital Research Institute (to A.D.M. and J.D.O), the Beatriu de Pinós postdoctoral programme of the Government of Catalonia’s Secretariat for Universities and Research of the Ministry of Economy and Knowledge (to J.D.O), and the Japan Society for the Promotion of Science 15H02421 and 18H04005 (S.K.). This work was partly funded by a Methuselah Foundation grant to J.P.M. To the Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) - Chile through the doctoral studentship number N°21170433 and the scholarship from MECESUP AUS 2003 to D.T.M. GenAge is funded by a Biotechnology and Biological Sciences Research Council (BB/R014949/1) grant to J.P.M. TMB is supported by BFU2017-86471-P (MINECO/FEDER, UE), Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d’Universitats i Recerca and CERCA Programme del Departament d’Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). C.F is supported by “La Caixa” doctoral fellowship program. EL is supported by CGL2017-82654-P (MINECO/FEDER, UE). The funding bodies played no role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.
Author contributions

ADM and JDO conceived of the project
JDO, ADM, YN, WCW, RK, JdC, JPM, and TMB contributed to the study design
JDO, DTM, MdM, MJM, LFK, and MCJ performed computational analyses
JDO, RK, and JT conducted flow cytometry
JDO, ADM, CF, JH, and EL conducted molecular lab work
ADM, GHP, ADF, JAH, and SK contributed samples
ADM, RK, and TMB provided computational and laboratory resources
JDO, ADM, MJM, JPM, and DTM wrote the manuscript with commentary from all authors

Acknowledgements

We thank R. Blanco Segura and M. M. Chavarria and staff from the Área de Conservación Guanacaste and Ministerio de Ambiente y Energía. From Kids Saving the Rainforest we thank the volunteers and veterinarians, especially Pia Martin, Carmen Soto, Jennifer Rice and Chip Braman. For assistance in the lab, we acknowledge Kelly Kries, Gwen Duytschaever, Jene Weatherhead, Laurie Kennedy, and Yiping Liu. We thank Aoife Doherty and Mengjia Li for exploratory analyses of selection in aging-related genes and acknowledge Patrick Minx and Kim Kyung for assistance with bioinformatics. We also thank Jessica Lynch and Hazel Byrne for comparative data. Thanks to R. Gregory in the Centre for Genomic Research and I.C. Smith in the Advanced Research Computing at the University of Liverpool for the access to the computing resources. To Dr. J.C Opazo for the insightful discussion.
Figure 1: Sector Santa Rosa (SSR) during wet (left) and dry (middle) seasons. Right: Map of sampling locations in Costa Rica. The two northern sites, SSR and Cañas, have tropical dry forest biomes, whereas the two southern sites, Quepos and Manuel Antonio are tropical wet forests. Photos - Amanda Melin; Drawing of white faced capuchin monkey - Alejandra Tejada-Martinez; Map: Eric Gaba – Wikimedia Commons user: Sting
Figure 2: Mapping percentages of sequencing reads from RNAlater preserved fDNA libraries prepared with FACS for A) all samples [Box-plot elements: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers], and B) individual libraries. C) Increase in mapping rate for RNAlater preserved samples. D) Relationship between mapped read duplication and number of cells with LOESS smoothing. The duplicate rate decreases sharply once a threshold of about 1,000 cells is reached.
Figure 3: Genes under positive selection in white-faced capuchin monkeys that are associated with longevity and brain development. Values from *Cebus capucinus* are used in place of *Cebus imitator*, given the recent taxonomic split.
Figure 4: Population subdivision in *Cebus imitator*. Left: Principal components of 13 fecal and 10 blood/tissue libraries from white faced capuchins. Individuals from northern and southern sites separate on PC 1. Low- and high-coverage *C. imitator* samples from Santa Rosa plot in the same cluster. Right: Maximum likelihood tree of 9 fecal and 10 blood/tissue libraries from *C. imitator* (samples with less than 0.5X coverage were excluded). Among the white-faced capuchin samples, individuals from northern (dry forest) and southern (wet forest) regions form the primary split; secondary splits reflect the individuals from different sites within regions. The short branch lengths of the outgroups are a result of only polymorphic positions within *C. imitator* being used to construct the tree.
Figure 5: A: Density plot of 1 Mb windows with a slide of 200 Kb in northern and southern populations. The distribution of windows from the northern population indicates lower heterozygosity than the southern distribution. The individuals from the southern population show consistently higher values. B: Long runs of homozygosity in the 5 largest scaffolds. Blue dots represent windows with depleted heterozygosity. The individuals with the longest runs of homozygosity come from the northern population. C: PSMC plot of effective population size over time.
Figure 6: Highly differentiated genes between wet and dry forest populations involved in diabetes, kidney function, and creatinine levels. 

A: Hudson’s $F_{ST}$ within windows of 20Kb with a 4Kb slide. Gene regions are in red, flanked by 500Kb (or length to beginning or end of scaffold) of sequence. X-axis values correspond to position along the scaffold. The dotted line indicates average $F_{ST}$ value across all windows ($F_{ST} = 0.178$), and the dashed line represents the top 0.5% of values ($F_{ST} = 0.797$). Vertical black lines indicate a non-synonymous SNP with an $Fst >= 0.750$, excluding BCAS3 (see Results). 

B: Heatmaps indicating the pattern of SNP variation within and surrounding highly divergent genes. SNVs within the genes are located under the red band and those within 200Kb of flanking region under the gray bands.
REFERENCES


38


77. Freitas AA, de Magalhães JP. A review and appraisal of the DNA damage theory of ageing. Mutat Res. 2011;728:12–22.


117. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and

