Mutationathon: towards standardization in estimates of pedigree-based germline mutation rates

Lucie A. Bergeron[1,*], Søren Besenbacher[2], Tychele N. Turner[3], Cyril J. Versoza[4], Richard J. Wang[5], Alivia Lee Price[1], Ellie Armstrong[6], Meritxell Riera[7], Jedidiah Carlson[8], Hwei-yen Chen[1], Matthew W. Hahn[5], Kelley Harris[9], April Snøfrid Lo Natalie M Kleppe[2], Elora H. López-Nandam[10], Priya Moorjani[11], Susanne P. Pfeifer[12], George P. Tiley[13], Anne D. Yoder[13], Guojie Zhang[1] and Mikkel H. Schierup[7,*]

1. Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen Ø, Denmark
2. Department of Molecular Medicine, Aarhus University, Brendstrupgårdsvej 21A, 8200 Aarhus N, Denmark
3. Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
4. Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85281 USA
5. Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN 47401 USA
6. Department of Biology, Stanford University, Stanford, CA, 94305 USA
7. Bioinformatics Research Centre, Aarhus University, C.F.Møllers Allé 8, 8000, Aarhus C, Denmark
8. Helix Inc., San Mateo, CA, 94401 USA
9. Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America, Computational Biology Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
10. California Academy of Sciences, 55 Music Concourse Drive, San Francisco, CA 94118 USA
11. Department of Molecular and Cell Biology, Center for Computational Biology, University of California, Berkeley, CA 94720 USA
12. Center for Evolution and Medicine, Center for Mechanisms of Evolution, School of Life Sciences, Arizona State University, Tempe, AZ 85281 USA
13. Department of Biology, Duke University, Durham, NC 27708 USA

Abstract

In the past decade, several studies have estimated the human per-generation germline mutation rate using large pedigrees. More recently, estimates for various non-human species have been published. However, methodological differences among studies in detecting germline mutations and estimating mutation rates make direct comparisons difficult. Here, we describe the many different steps involved in estimating pedigree-based mutation rates, including sampling, sequencing, mapping, variant calling, filtering, and how to appropriately account for false-positive and false-negative rates. For each step, we review the different methods and parameter choices that have been used in the recent literature. Additionally, we present the results from a "Mutationathon", a competition organized among five research labs to compare germline mutation rate estimates for a single pedigree of rhesus macaques. We report almost a two-fold variation in the final estimated rate among groups using different post-alignment processing, calling, and filtering criteria and provide details into the sources of variation across studies. Though the difference among estimates is not statistically significant, this discrepancy emphasizes the need for standardized methods in mutation rate estimations and the difficulty in comparing rates from different studies. Finally, this work aims to provide guidelines for computational and statistical benchmarks for future studies interested in identifying germline mutations from pedigrees.

Introduction

Germline mutations are the source of most genetic diseases and provide the raw material for evolution. Thus, it is crucial to accurately estimate the frequency at which mutations occur in order to better understand the course of evolutionary events. The development of high throughput next-generation sequencing offers the opportunity to directly estimate the germline mutation rate over a single generation, based on a whole-genome comparison of pedigree samples (mother, father, and offspring), without requiring assumptions about generation times or fossil calibrations (Tiley et al., 2020). Pedigree sequencing provides multiple pieces of information in addition to an overall mutation rate. For instance, the genomic locations, the spectrum of mutation types (e.g. transition or transversion), and the nucleotide context of all mutations can easily be gleaned. Furthermore, pedigree sequencing enables researchers to identify the parental origin of the mutations--that is, whether the

mutation arose in the maternal or paternal germline. Finally, using pedigrees means that researchers often have precise information about the age of the parents at the time of reproduction, and comparing several trios (i.e. three related individuals: mother, father, and offspring) at different parental ages can tell us about the effect of parental age on the total number of transmitted mutations, their location, and their spectrum. Thus, there has been a growing interest in applying this method to address medical and evolutionary questions.

The first estimate of the human germline mutation rate using pedigrees was published more than ten years ago (Roach et al., 2010). Four years later, the first pedigree-based mutation rate for a non-human primate, the chimpanzee, was estimated (Venn et al., 2014). Today, at least 20 vertebrate species have mutation rates estimated by pedigree sequencing (Table 1), with half added in the past two years. Each study differs in the number of trios, the sequencing technology and depth, the ages of individuals included, and the bioinformatics pipelines used to analyze the data (see Table 1 and Supplementary Table 1). Thus, reported variation in mutation rates among studies might result from a combination of biological and methodological factors. With an increasing number of studies being published, an examination of the differences among studies and suggestions for standards that will minimize differences caused by methodological discrepancies are warranted.

Table 1 – Vertebrate species with a direct estimate of the mutation rate using a pedigree approach. The list of species includes ten primates, five non-primate mammals, one bird, and four fish (see Supplementary Table 1 for differences in study design and methodology).

Species	**Mutation rate per site per generation : $\mu \times 10^{-8}$**	**Number of trios**	**Parental age ***	**Reference**
Orangutan (Pongo abelii)	1.66	1	♂: 31.00 and ♀: 15.00	(Besenbacher et al., 2019)

Human (*Homo sapiens*)	1.17	1 (CEU)	unspecified	(Conrad et al., 2011)
	0.97	1 (YRI)	unspecified	(Conrad et al., 2011)
	1.20	78	♂: 29.10 and ♀: 26.50	(Kong et al., 2012)
	1.20	269	unspecified	(Francioli et al., 2015)
	1.28	13	♂: 29.80	(Rahbari et al., 2016)
	1.05	719	♂: 33.40	(Wong et al., 2016)
	1.29	1550	unspecified	(Jónsson et al., 2017)
	1.28	150	~ 27.70	(Maretty et al., 2017)
	1.30	516	♂: 33.40	(Turner et al., 2017)
	1.10	593	♂: 29.10 and ♀: 26.00	(Sasani et al., 2019)
	1.22	1449	♂: 29.70 and ♀: 26.90	(Kessler et al., 2020)
Chimpanzee (*Pan troglodytes*)	1.20	6	♂: 18.90 and ♀: 15.00	(Venn et al., 2014)
	1.48	1	♂: 24.00 and ♀: 24.00	(Tatsumoto et al., 2017)
	1.26	7	♂: 19.30 and ♀: 15.90	(Besenbacher et al., 2019)
Gorilla (*Gorilla gorilla*)	1.13	2	♂: 14.50 and ♀: 20.50	(Besenbacher et al., 2019)
Baboon (*Papio anubis*)	0.57	12	♂: 10.70 and ♀: 10.20	(Wu et al., 2020)
Rhesus macaque (*Macaca mulatta*)	0.58	14	♂: 7.80 and ♀: 7.10	(Wang et al., 2020)
	0.77	19	♂: 12.40 and ♀: 8.40	(Bergeron et al., 2021)
Green monkey (*Chlorocebus sabaeus*)	0.94	3	♂: 8.70 and ♀: 4.70	(Pfeifer, 2017)
Owl monkey (*Aotus nancymaae*)	0.81	14	♂: 6.60 and ♀: 6.50	(Thomas et al., 2018)

Marmoset (*Callithrix jacchus*)	0.43	1	~ 2.80	(Yang et al., 2021)
Gray mouse lemur (*Microcebus murinus*)	1.52	2	♂: 4.55 and ♀: 1.45	(Campbell et al., 2021)
Mouse (*Mus musculus*)	0.57 0.39	8 15	unspecified ~ 0.47	(Milholland et al., 2017) (Lindsay et al., 2019)
Cattle (*Bos taurus*)	1.17	5	unspecified	(Harland et al., 2017)
Wolf (*Canis lupus*)	0.45	4	♂: 4.00 and ♀: 2.25	(Koch et al., 2019)
Domestic cat (*Felis catus*)	0.86	11	♂: 4.70 and ♀: 2.90	(Wang et al., 2021b)
Platypus (*Ornithorhynchus anatinus*)	0.70	2	unspecified	(Martin et al., 2018)
Collared flycatcher (*Ficedula albicollis*)	0.46	7	unspecified	(Smeds et al., 2016)
Herring (*Clupea harengus*)	0.20	12	unspecified	(Feng et al., 2017)
Cichlid (*Astatotilapia calliptera*, *Aulonocara stuartgranti* and *Lethrinops lethrinus*)	0.35	9	unspecified	(Malinsky et al., 2018)

* Depending on the study, the parental ages are reported as average paternal age (♂), average maternal age (♀), average parental age (~), or unspecified.

The key principle of the pedigree-based approach is to detect *de novo* mutations (DNMs) present in a heterozygous state in an offspring that are absent from its parents' genomes (Figure 1). A per-site per-generation mutation rate can be inferred by dividing the number of DNMs by the number of sites in the genome that mutations could possibly be identified in

(and accounting for the diploid length of the genome, as mutations can be transmitted by both the mother and father). As mutations are rare events, detecting all the true DNMs (or having a high sensitivity) while avoiding errors (or increasing precision) from a single generation remains challenging. False-positive calls (sites incorrectly detected as DNMs) can be caused by sequencing errors, errors introduced by read-mapping and genotyping steps, or somatic mutations in the offspring. Numerous filters are thus often applied on the variant sites to increase the precision of the candidate DNMs detection. However, filters that are too conservative can also discard true DNMs, reducing the sensitivity by increasing the rate of false-negative calls (true DNMs not detected). Therefore, a balance should be found between precision and sensitivity--a goal that has led to the development of multiple different methods to estimate germline mutation rates from pedigree samples.

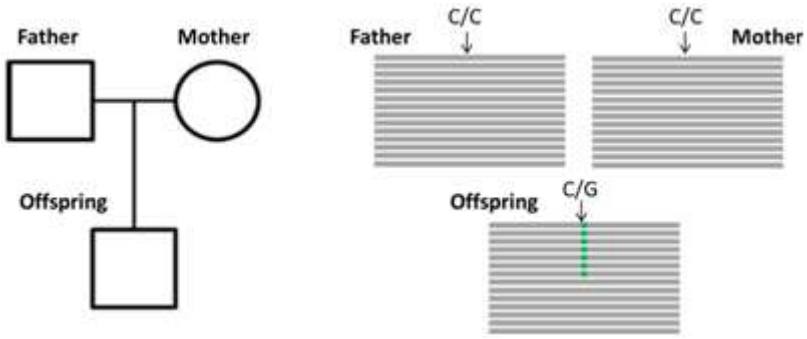

Figure 1 – Detection of a *de novo* mutation (DNM) in a trio sample (mother, father, and offspring). Potential candidates for DNMs are sites where approximately half of the reads (indicated as grey bars) from the offspring have a variant (indicated in green) that is absent from the parental reads.

In this study, we aim to define what we consider to be the state-of-the-art in pedigree-based germline mutation rate estimation, to discuss the pros and cons of each methodological step, and to summarize best practices that should be used when calling germline mutations. We review several recently published methods that estimate germline mutation rates from pedigree samples. In parallel, we set up a competition--the "Mutationathon"--among five research groups to explore the effect of different methodologies on mutation rate estimates. Using a common genomic dataset consisting of a pedigree of the rhesus macaque (*Macaca mulatta*; Bergeron et al., 2021), each group estimated the number of candidate DNMs (validated by PCR amplification and Sanger resequencing) and a germline mutation rate. An examination of the estimated rates produced by different groups not only highlighted the

choices that can be made in estimating per-generation mutation rates, but it also provided us with an opportunity to characterize the impact of these choices on the systematic differences in estimated rates, which in turn yielded important insights into the parameters that could reduce the occurrence of false-positive calls.

Results

Comparison of methods

The overall pipeline from high throughput next-generation sequencing data to an estimated mutation rate is similar across all studies listed in Table 1. It includes five steps (Figure 2):

1. sampling and whole-genome sequencing of at least one trio or extended pedigrees that also include a third-generation (useful for validation of putative DNMs in the offspring),

2. alignment of reads to a reference genome and post-processing of alignments,

3. variant-calling to infer genotypes or genotype likelihoods for all individuals,

4. detection of DNMs via filtering of candidates (including an assessment of the false discovery rate), and finally

5. the estimation of a per-generation mutation rate accounting for the length of the accessible genome (including an assessment of the false-negative rate).

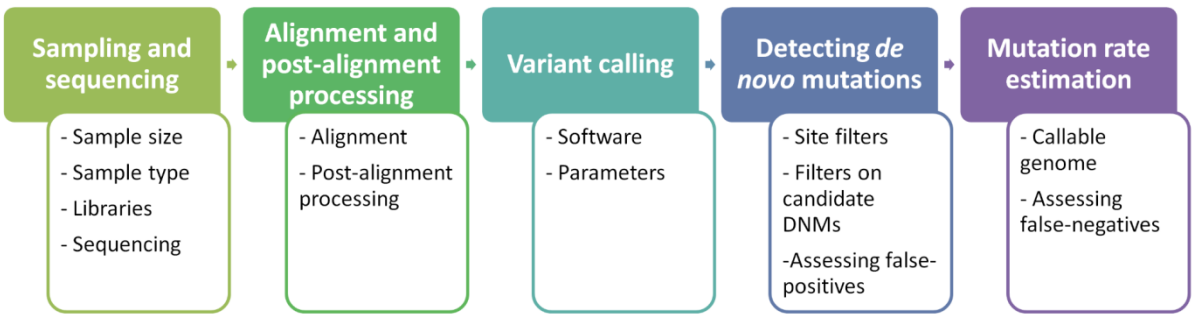

Figure 2 – Flow of the main steps to call *de novo* mutations (DNMs) from pedigree samples. Each step lists the various choices in study design and methodology that might impact mutation rate estimates.

Step 1 - Sampling and sequencing

Sample size. Pedigree-based study designs can vary significantly, from those that include only one trio (e.g., Besenbacher et al., 2019) to those that include thousands of trios (e.g., Halldorsson et al., 2019). The first study to estimate a pedigree-based human mutation rate used only two trios and estimated a mutation rate of 1.1×10^{-8} per site per generation (Roach et al., 2010), which is within the overall variation reported across studies with larger sample sizes (Table 1). Larger sample sizes reduce uncertainty in the average mutation rate for a species and offer more statistical power for the exploration of various parameters such as the parental age effect, the contribution of each parent to the total number of DNMs, the presence of mosaicism and shared mutations when including siblings, or the distribution of mutations across genomes. Moreover, multiple generation pedigrees, also referred to as extended trios, can be used to validate true DNMs and adjust quality filters by studying transmission to a third generation. Therefore, whenever possible, multiple trios should be analyzed and more than one generation should be included. Finally, the age of the parents at the time of reproduction is required for estimating the per-year mutation rate from the per-generation rates directly measured in the trios. In some studies, the age of the parents at conception is not available, and instead, the mean age of reproduction is used for the estimation of the per-year mutation rate. While useful, this approximation can lead to biased results if the age of the parents at conception was much older or much younger compared to the mean age in the population. Thus, when possible, the information on the age of each parent at the time of conception should be collected as it is essential for the interpretation of results and to help understand parental age effects on mutation rate.

Sample type. The most commonly used sample types are somatic tissues such as whole blood, muscle, or liver, which generally produce a high quantity of DNA with long fragment sizes and allow for high-coverage sequencing. The duration and temperature of storage can affect the quality of the extracted DNA and increase the rate of sequencing errors. Thus, to minimize DNA damage during storage, DNA is typically kept in TE (tris-EDTA) buffer. Moreover, it is advised to store DNA at -80°C for long-term storage (months to years) and in liquid nitrogen at -164°C for decades (Baust, 2008; Straube and Juen, 2013). Other materials such as buccal swabs or fur can be considered, but they can be technically challenging. For instance, as part of a recent study on rhesus macaques (Bergeron et al., 2021), DNA was extracted from hair samples and sequenced at 95X coverage, yet, due to the fragmentation,

only 38% of the reads were mappable to the reference genome. After variant calling, the average depth of usable reads was 6X, with only 10% of sites covered by more than 10 reads. To reduce the number of false-positive calls caused by somatic mutations, it is best to avoid tissues with an accumulation of such mutations, such as skin. In this regard, blood is often the preferred tissue: as many different tissues contribute cells to the blood, the hope is that a somatic mutation in any one of them will not be mistaken for a DNM. However, in rare cases, mainly in older individuals, clonal hematopoiesis can lead to high frequency somatic mutations in the blood. Sequencing more than one type of tissue, when feasible, should be considered in order to reduce the potential for mistaking somatic mutations as candidate DNMs.

Libraries. After DNA extraction, genomic library preparation is another step that can introduce sequencing errors. Most studies have used Illumina sequencing platforms, yet, even for a single technology, there are different library preparation protocols available. PCR amplification is commonly used to increase the quantity of DNA, but this can generate artifacts caused by the introduction of sequence errors (PCR errors) or by the over-amplification of some reads (PCR bias) (Acinas et al., 2005). Thus, for samples yielding a sufficient amount of DNA, PCR-free libraries that do not involve amplification prior to cluster generation are preferable. Moreover, as different library preparation methods can result in different amplification biases (Ross et al., 2013; Wingett, 2017), utilizing different types of library preparations may be advisable to reduce the sources of error.

Sequencing. All Illumina sequencing platforms use similar sequencing chemistry (sequencing-by-synthesis) and mainly differ in running speed and throughput. Another equivalent technology, used in two studies (Bergeron et al., 2021; Roach et al., 2010), is BGISEQ-500, combining DNA nanoball nanoarrays with polymerase-based stepwise sequencing (Mak et al., 2017) and showing similar performances to Illumina on data quality (Chen et al., 2019; Patch et al., 2018). Another study used 10x Genomics linked reads, which can help phase maternal and paternal mutations (Campbell et al., 2021). However, it remains unclear if alternative library preparation and sequencing platforms introduce additional biases compared to standard Illumina protocols. Most pedigree-based studies of germline mutations have sequenced each individual to a depth between 30X and 50X (Besenbacher et al., 2019; Campbell et al., 2021; Jónsson et al., 2017; Kessler et al., 2020; Malinsky et al., 2018; Milholland et al., 2017; Sasani et al., 2019; Smeds et al., 2016; Thomas et al., 2018; Turner et

al., 2017; Wang et al., 2020; Wu et al., 2020), three studies sequenced at a higher depth of ~ 80X (Bergeron et al., 2021; Maretty et al., 2017) and 150X (Tatsumoto et al., 2017), while six studies sequenced at a depth lower than 25X on average (Harland et al., 2017; Koch et al., 2019; Lindsay et al., 2019; Martin et al., 2018; Pfeifer, 2017; Rahbari et al., 2016). A minimum coverage of 15X has been advised to call single nucleotide polymorphisms (SNPs) accurately (Fumagalli et al., 2013). Yet, this depth might not be sufficient to call germline mutations, since it might be hard to distinguish genuine germline mutations from somatic mutations that are present in a substantial fraction of cells. Furthermore, with low coverage the probability of calling a parent homozygous for the reference allele, when they are actually heterozygous, becomes non-negligible at the genome-wide level. For example, the binomial probability of not observing a read with one of the alleles in a heterozygote with 15X coverage is $0.5^{15} = 3.05 \times 10^{-5}$, which will happen by chance around 30 times in a genome with 1 million heterozygous positions. Likewise, based on the binomial distribution, the probability that a somatic mutation present in 10% of cells is seen in more than 30% of reads is 0.0113 with 20X coverage but falls to 0.0004 with 35X coverage. Thus, it is advised to aim for a minimum of 35X as a rule of thumb.

Step 2 - Alignment and post-alignment processing

Alignment. To find DNMs, we must first find where in the genome each of the short sequencing reads comes from. The Burrows-Wheeler Aligner (BWA; Li and Durbin, 2009) is an algorithm developed to map short reads (50-250 bp) to a reference genome and has been used in the majority of studies on direct mutation rate estimation (Bergeron et al., 2021; Besenbacher et al., 2019; Harland et al., 2017; Jónsson et al., 2017; Kessler et al., 2020; Koch et al., 2019; Malinsky et al., 2018; Maretty et al., 2017; Milholland et al., 2017; Pfeifer, 2017; Sasani et al., 2019; Smeds et al., 2016; Tatsumoto et al., 2017; Thomas et al., 2018; Turner et al., 2017; Wang et al., 2021b, 2020; Wu et al., 2020). In particular, the BWA-MEM algorithm is fast, accurate, and can be implemented with an insert size option to improve the matching of paired reads. Several aspects of the study organism and study design can have detrimental effects on read mapping. Some studies reported a trimming step to remove adapter sequences and poor-quality reads--those with a high proportion of unknown ("N") bases or low quality-score bases (Bergeron et al., 2021; Maretty et al., 2017; Tatsumoto et al., 2017; Wu et al., 2020). However, trimming might not be necessary as some mapping software will soft-clip (or mask) the adaptors, while low-quality reads can be removed during

the variant calling step. The quality of the reference genome can play an important role in obtaining a large proportion of reads with high mapping scores. In the case of a poor or non-existent reference genome, using the reference genome of a phylogenetically related species is an option, but this could make the downstream analysis more complex (Prasad et al., 2021). Moreover, BWA was designed to map low-divergence sequences, so that using a related species, or even a closely related individual in the same species when heterozygosity is high, could impact the mapping. Finally, low-complexity regions and repetitive sequences such as dinucleotide tandem repeats can be problematic for read-mapping. Standards have been proposed for human genome analysis and can be followed for germline mutation rate calling in species with comparable heterozygosity (for details, see Regier et al., 2018).

Post-alignment processing. To correct for possible misalignment of sequencing reads to the reference genome, post-alignment quality control is necessary. This step often includes base-quality score recalibration (BQSR), removing of duplicate reads, and realignment around indels. BQSR corrects for any bias in the base quality score assigned by the sequencer by utilizing information from a set of known variants for the studied species. When such a dataset is not available, as in many non-human species, the Best Practices of the Genome Analysis ToolKit (GATK) software from the Broad Institute advises to proceed first with variant calling in all available samples and subsequently using the best quality variants to recalibrate the base quality scores (GATK team, 2021). If multiple generations are available, high-quality variants fully transmitted across generations can be used for BQSR (Wu et al., 2020). However, some studies have ignored this step due to the circularity of this method and its computational expense, as variants will be called twice (Bergeron et al., 2021; Thomas et al., 2018; Wang et al., 2020). A comparative study presented a difference of less than 0.1 % between the total variant sites called with and without recalibration (Li, 2014), and this difference was even lower for high-coverage (40X) sequencing (Tian et al., 2016); however, this step is still advised to increase the quality of variant calling (Li, 2020). Duplicates, identical reads due to amplification (PCR duplicates) or sequencing clusters (optical duplicates), can increase false-positive calls and erroneously inflate sequencing coverage. Therefore, duplicates should be marked or removed even for sequences from PCR-free libraries. Reads terminating with indels are more likely to be misaligned; thus, depending on the variant caller used, realignment around indels may be advised to correct for this artifact. Specifically, realignment around indels is required when calling variants with non-haplotype-aware callers (such as GATK's UnifiedGenotyper), but is not necessary with haplotype-

aware variant callers (such as GATK HaplotypeCaller (Poplin et al., 2018), Platypus (Rimmer et al., 2014), or FreeBayes (Garrison and Marth, 2012)). From GATK release 3.6 onward, the realigned reads around indels can be outputted during the variant calling step. Alternatively, BWA alignments can be used to construct a variation-aware graph with GraphTyper (Eggertsson et al., 2017), including known polymorphisms and newly genotyped variants. Thereby, reads are re-aligned to the graph, reducing reference bias and improving read alignment near indels (Eggertsson et al., 2017) and structural variants (Eggertsson et al., 2019). Finally, other quality controls can be applied after mapping, such as removing reads mapping to multiple locations, as they could map with a good mapping quality in two or more locations and be ignored by further quality filters. However, the overall impact of many of these filters, such as BQSR and realignment around indels, on the final set of DNMs has not yet been studied.

Step 3 - Variant calling

Software. Different algorithms have been shown to perform similarly well in calling nucleotide variants (Li, 2014). GATK (Auwera and O'Connor, 2020) is widely used among studies that call germline DNMs (Bergeron et al., 2021; Besenbacher et al., 2019; Campbell et al., 2021; Feng et al., 2017; Harland et al., 2017; Jónsson et al., 2017; Koch et al., 2019; Malinsky et al., 2018; Maretty et al., 2017; Milholland et al., 2017; Pfeifer, 2017; Sasani et al., 2019; Smeds et al., 2016; Tatsumoto et al., 2017; Thomas et al., 2018; Turner et al., 2017; Wang et al., 2021b, 2020; Wong et al., 2016; Wu et al., 2020). Other commonly used variant callers are GraphTyper (Eggertsson et al., 2017; e.g., utilized by Beyter et al., 2021; Halldorsson et al., 2019; Jónsson et al., 2021, 2018) and FreeBayes (Garrison and Marth, 2012; e.g., utilized by Turner et al., 2017). Using more than one variant caller can increase confidence in the SNP set but can become computationally expensive (Turner et al., 2017).

Parameters. Even within the same variant caller, different methods can be used (see Supplementary Table 1). For instance, in GATK v3, three strategies are available: 1. per-sample variant calling, 2. batch calling, in which samples are analyzed separately and concatenated for downstream analysis, and 3. joint calling, in which variants are called simultaneously across all samples (with the UnifiedGenotyper command). In GATK v4, the new recommendation is to first call variants for each sample separately (HaplotypeCaller in ERC mode), and then combine all the samples (GenomicsDBImport) to jointly genotype them (GenotypeGVCFs)--thus, the initial identification of variant sites is separable from the

assignment of genotypes to each individual. UnifiedGenotyper and HaplotypeCaller should have a similar ability to detect SNPs, but differences in variant sets have been observed (Lescai et al., 2014). Moreover, the GATK HaplotypeCaller ERC mode has two options: the BP_RESOLUTION option provides records for every single site in the genome, even non-variant sites, while the GVCF option groups the non-variant sites into a block of record. This variant-calling step is computationally expensive, especially if variants are called in BP-RESOLUTION mode, but it can be useful to determine the part of the genome in which there is full power to detect mutations. It is still unclear which strategy should be prioritized; thus, it is advised to report the method used and any additional options that have been implemented. The default settings of GATK applied during variant calling should also be kept in mind. For instance, the heterozygosity prior is by default at 0.001, which could have an impact when analyzing species with much higher or much lower heterozygosity, though the effect of this prior has not been evaluated in the context of mutation rate studies.

Step 4 - Detecting *de novo* mutations

Site filters. GATK's Best Practices (Auwera and O'Connor, 2020) advise a Variant Quality Score Recalibration (VQSR) step to ensure that genotypes are correctly called. However, this tool is not suitable for DNMs as it would remove many rare variants; instead, hard-filtering should be applied. GATK provides some general recommendations for these filters, warning that these should be a starting point and filters may need to be adjusted depending on the callset or the species studied (GATK team, 2020). The currently advised hard filter criteria for germline short variant discovery are: QD < 2.0; MQ < 40.0; FS > 60.0; SOR > 3.0; MQRankSum < -12.5; ReadPosRankSum < -8.0. These parameters take into account the quality of a call at a given site (QD), the mapping quality (MQ), the strand bias (FS and SOR), the mapping quality bias between reference and alternative allele (MQRankSum), the position bias within reads (ReadPosRankSum); see Supplementary Table 2 for details on each filter). Although some studies followed these best practices (Jónsson et al., 2017; Wu et al., 2020), others implemented only a subset of filters (e.g., three studies reported the GATK filters without SOR > 3.0 (Koch et al., 2019; Thomas et al., 2018; Wang et al., 2020) and Besenbacher et al. (2019) kept only four parameters -- FS, ReadPosRankSum, BaseQualityRankSum, and MQRankSum -- as they are calculated based on statistical tests following a known distribution) or readjusted the filtering thresholds based on previous results (e.g., Koch et al. (2019) changed the ReadPosRankSum threshold from -8 to 15 while

Bergeron et al. (2021) changed the site filters by first using the advised parameters and then adjusting them to reduce the apparent false-positive calls). Several of the earlier studies implemented a different suite of site filters altogether (Pfeifer, 2017; Smeds et al., 2016; Tatsumoto et al., 2017). Given this plethora of choices, we suggest that reporting filter details should be common practice to improve the comparability of mutation rate estimates. Another site filter is the Phred-scaled probability that a certain site is polymorphic in one or more individuals (QUAL), which has been used in some studies (Harland et al., 2017; Pfeifer, 2017; Wu et al., 2020).

Filters of candidate DNMs. From pedigrees, germline mutations are detected as "Mendelian violations" where at least one of the alleles observed in the offspring is absent from both of its parents. Most mutation rate studies restrict Mendelian violations to sites where both parents are homozygous for the reference allele (HomRef; 0/0) and the offspring is heterozygous (Het; 0/1 or 1/0) (Bergeron et al., 2021; Besenbacher et al., 2019; Jónsson et al., 2017; Koch et al., 2019; Pfeifer, 2017; Smeds et al., 2016; Thomas et al., 2018; Wang et al., 2020; Wu et al., 2020). Other combinations of genotypes could also be caused by germline mutations such as parents homozygous for the alternative allele (HomAlt; 1/1) with heterozygous offspring (0/1 or 1/0), or one parent HomRef (0/0) and the other HomAlt (1/1) with an offspring either HomRef (0/0) or HomAlt (1/1). These sites are usually filtered out and assumed to represent a small portion of the genome to avoid the added uncertainty associated with these genotypes (Wang et al., 2021a). However, before excluding these sites, researchers should note that their expected frequency increases with the level of heterozygosity of the species studied and the phylogenetic distance to the reference genome used for mapping. For a phylogenetic distance to the reference genome of 2%, ~1 in 50 true DNMs is expected to occur in a background where both parents are homozygous for the alternative allele (1/1). After selecting the final set of Mendelian violations, several filters are applied to ensure the genotypes of each individual are of high quality and to reduce false-positive calls. The individual filters and thresholds used vary substantially between studies (see Supplementary Table 3), but generally include a depth filter (i.e., the number of reads for each individual at a particular site), a genotype quality filter (i.e., the Phred-scaled confidence of the assigned genotype), as well as a filter on the allelic depth (i.e., the number of reads supporting the alternative allele and the reference allele).

Sites with low read depth (DP) are prone to exhibit Mendelian violations due to sequencing and genotyping errors, while positions with particularly high depth could indicate a misalignment of reads in low complexity or paralogous regions. As each study analyzed pedigrees sequenced at various depths, different cutoffs were chosen for this filter, some more permissive than others. Some studies only set a minimum DP of approximately 10 reads (e.g., Jónsson et al., 2017; Pfeifer, 2017; Sasani et al., 2019), while other higher coverage studies were able to set more conservative minimum and maximum thresholds, varying from a minimum of 10 to 20 to a maximum of 60 to 150 (e.g., Maretty et al., 2017: DP < 10 and DP > 150; Thomas et al., 2018: DP < 20 and DP > 60; Wang et al., 2020: DP < 20 and DP > 60). Another approach is to use a relative depth threshold for each individual (e.g., $depth_{individual} \pm 3\sigma$, with σ being the standard deviation around the average depth (Tatsumoto et al., 2017), or a maximum threshold of $2 \times depth_{individual}$ (Besenbacher et al., 2019)) or, when all individuals were sequenced at a similar depth, an relative depth per trio (e.g., a DP filter of $0.5 \times depth_{trio}$ and $2 \times depth_{trio}$ (Bergeron et al., 2021)). Alternatively, Rahbari et al. (2016) and Wu et al. (2020) tested if the depth at each site followed a Poisson distribution under the null hypothesis that lambda was $depth_{individual}$, and filtered away sites where at least one individual of the trio had a p-value higher than 2×10^{-4}.

To correct for genotyping errors, two parameters from the output variant-calling file can be used: the Phred-scaled likelihood of the genotype (PL) and the genotype quality (GQ). The most likely genotype has a PL of 0, while the least likely genotype has the highest PL value. GQ is the difference between the $PL_{2nd\ most\ likely}$ and $PL_{1st\ most\ likely}$, with a maximum reported of 99. Applied GQ thresholds vary between 20 (Jónsson et al., 2017; Sasani et al., 2019) and 70 (Wang et al., 2021b, 2020). Instead of using GQ, some studies used the difference between $PL_{2nd\ most\ likely}$ and $PL_{1st\ most\ likely}$, which is not limited to a maximum of 99, and applied more conservative criteria for the offspring heterozygous genotype than for the homozygous parents (Maretty et al., 2017: homozygous $PL_{2nd\ most\ likely} - PL_{1st\ most\ likely} < 80$, heterozygous $PL_{2nd\ most\ likely} - PL_{1st\ most\ likely} < 250$; Tatsumoto et al., 2017: homozygous $PL_{2nd\ most\ likely} - PL_{1st\ most\ likely} < 100$, heterozygous $PL_{2nd\ most\ likely} - PL_{1st\ most\ likely} < 200$).

Variants can also be filtered using allelic depth: the number of reads supporting the reference allele and the alternative allele. To ensure the homozygosity of the parents, some studies filter away sites where alternative alleles are present in the parents' reads. AD refers to the number of reads supporting the alternative allele, with previously utilized thresholds include AD > 0

(Besenbacher et al., 2019; Harland et al., 2017; Koch et al., 2019; Pfeifer, 2017; Sasani et al., 2019; Smeds et al., 2016; Wang et al., 2021b), AD > 1 (Jónsson et al., 2017; Wang et al., 2020), or AD > 4 (Maretty et al., 2017). Even more conservative, one study used a lowQ AD2 > 1, i.e. the number of alternative alleles in the low-quality reads (not used for variant calling) should not exceed 1 (Besenbacher et al., 2019).

Allelic depth is also used to calculate the allelic balance (AB): the proportion of reads supporting the alternative allele relative to the total depth at this position. In the case of a DNM, the offspring should have approximately 50% of its reads supporting each allele. Purely somatic mutations are expected to cause only a small fraction of reads to carry an alternate allele, though this fraction can be different for mutations occurring early in the zygote stage of the offspring and leading to germline mosaicism. A previous large-scale analysis of human pedigrees recovered a bi-modal allelic balance distribution of Mendelian violations in the offspring before applying an AB filter, with a peak around 50% interpreted as DNMs, and another peak around 20% likely corresponding to somatic mutations (Besenbacher et al., 2015), mismapping errors, or sample contamination (Karczewski et al., 2019). Thus, careful filtering on AB is required to avoid false positives. Thresholds used for the AB filter vary between a minimum of 20% (Pfeifer, 2017) to 40% (Thomas et al., 2018), and a maximum, when applied, of 60% (Thomas et al., 2018) to 75% (Jónsson et al., 2017). Instead of a hard cutoff, one study used a binomial test on the allelic balance under the null hypothesis of a 0.5 frequency, removing positions with a p-value lower than 0.05 (Wu et al., 2020).

Additional filters can be used, for instance, to remove candidate DNMs present in individuals other than the focal offspring, including siblings (Pfeifer, 2017; Smeds et al., 2016), only unrelated individuals in the same dataset (Bergeron et al., 2021; Besenbacher et al., 2019; Campbell et al., 2021; Thomas et al., 2018; Wu et al., 2020) or polymorphism datasets of the same species (Pfeifer, 2017; Smeds et al., 2016; Wu et al., 2020). This filter is based on the idea that the chance of getting a DNM at a position already being polymorphic is very low unless there is very high heterozygosity, thus guarding against the possibility that a heterozygous site was missed in the parents. Filters can also be applied to the distance between mutations, again assuming that the probability of having two mutations close to each other is low. For instance, in some studies, candidate DNMs were removed if four or more candidates were located in a 200 base-pairs window (Koch et al., 2019), or two candidates

were less than 10 base-pairs (Tatsumoto et al., 2017) or 100 base-pairs apart (Wu et al., 2020) from each other. However, the underlying assumptions for these filters are not always fulfilled. Indeed, recurrent mutations can occur, especially at CpG locations (Acuna-Hidalgo et al., 2016; Ségurel et al., 2014), and there is evidence of nonrandom clustering of mutations (Brandler et al., 2016; Turner et al., 2016). Finally, some studies removed DNM candidates located in the locus control region (LCR; Sasani et al., 2019) or repetitive regions of the genome (Pfeifer, 2017) which are prone to mismapping.

Assessing false-positives. After choosing each filter according to the dataset, a total number of candidate DNMs per offspring is found. Yet, as stringent as the filters can be, there are still chances for false positives (FPs) to be introduced in the final set of DNMs. Even though there is no perfect method to correct the false-positive calls, this issue should be addressed.

One of the most straightforward methods to validate DNMs is by PCR amplification followed by resequencing such as Sanger sequencing, to ensure the genotype of each individual of the trio (Bergeron et al., 2021; Koch et al., 2019; Maretty et al., 2017; Tatsumoto et al., 2017; Wu et al., 2020). However, this PCR amplification and resequencing method can be challenging. In addition to the cost, designing primers for the region of the candidate DNMs can be difficult, especially for candidates located in repeat regions. Furthermore, most Sanger resequencing is aimed at validating the heterozygous state of the offspring, not the homozygous state of the parents. If all candidate DNMs are successfully validated, the false positives can be removed from the set of candidate DNMs. However, it is often the case that we cannot check every candidate DNM. In these cases, it is common to estimate the false discovery rate (FDR) from a subset of candidates that can be checked. The FDR can be estimated as:

$$FDR = \frac{PCR_{failed}}{PCR_{validated} + PCR_{failed}},$$

with $PCR_{validated}$ being the number of candidate DNMs successfully amplified and passing the resequencing validation and PCR_{failed} being the number of candidate DNMs successfully amplified but failing the resequencing validation. We can then adjust the total number (nb) of DNMs in the entire dataset by using the following relationship:

$$nb_{candidate\ DNMs\ corrected} = nb_{candidate\ DNMs} \times (1 - FDR),$$

where $nb_{candidate\ DNMs\ corrected}$ is the updated number of DNMs in the dataset. Of note, some studies refer to a false positive rate instead of the FDR (eg. Bergeron et al., 2021; Jónsson et al., 2017; Wang et al., 2020), yet, it also refers to the ratio of false-positive calls on the total number of candidates (i.e. true positives and false positives).

A second method to check candidate DNMs is manual curation, using visualization software such as the Integrative Genome Viewer (IGV; Robinson et al., 2011). By comparing the read mappings of the parents and their offspring at candidate DNMs, false-positive calls can be detected. Estimates of the false discovery rate using this approach have varied widely depending on the study design, from 91% (Pfeifer, 2017) at low coverage to 35% (Smeds et al., 2016) at medium coverage to 11% at high coverage (Bergeron et al., 2021). Further work is needed to ensure that manual curation is consistent when applied by different researchers working in different systems.

A third method to estimate the false discovery rate, based on deviations from the expected 50% transmission rate of DNMs to the next generation, can be used if an extended pedigree is available. With this method, Wu et al. (2020) estimated a false discovery rate of 18%. However, such a deviation from 50% can arise from the expected variance of a binomial distribution, especially if the number of mutations is small. Moreover, clusters of mutations could increase this variance if linked mutations are passed on together to the next generation, especially if the number of trios is small. When this method is used, transmission should be clearly defined as it can be when the grandchild has been genotyped as heterozygote with the mutant allele, or alternatively when at least a few reads contain the mutant allele. Jónsson et al. (2017) used multiple individuals and haplotype sharing to assess the consistent segregation of DNM allele in the next generation.

A fourth method of estimating the false discovery rate takes advantage of monozygotic twins. Germline mutations transmitted from parents to monozygotic twins are expected to be present in both twins, as they are derived from the same zygote. Jónsson et al. (2017) exploited the discordance between candidate DNMs in monozygotic twins to derive the false discovery rate (3%). This estimate is an upper bound because discordance between monozygotic twins is a combination of post-zygotic mutations and false-positive calls. However, the authors analyzed a unique dataset of 91 human trios with monozygotic twins -- data that will be hard to obtain in most species.

Step 5 – Mutation rate estimation

To calculate a per-site per-generation mutation rate, the total number of candidate DNMs (corrected for false positives), should be divided by the number of sites in the genome with full detection power. The denominator takes into account the callable genome (CG) - sites where mutations could have been detected, and the false-negative rate (FNR) - the rate at which actual DNMs have been missed by the pipeline that has been applied to this point. Assuming that the rate of mutation is similar in the remaining part of the genome, the mutation rate per-site per-generation μ of a diploid species can be estimated as:

$$\mu = \frac{nb_{candidate\ DNMs} \times (1-FDR)}{2 \times CG \times (1-FNR)}.$$

Callable genome. Different methods have been used to estimate the CG, the number of sites where a DNM would have been detected if it was there (Supplementary Table 1). Many studies used the strict individual filters applied during the detection of candidate DNMs, including all sites where the parents were homozygotes for the reference allele and each individual met the DP, GQ, and any other filters. However, the set of filters and input files used to infer CG differ between studies (Supplementary Table 1) and, consequently, estimates vary widely from CG representing 45% (Tatsumoto et al., 2017) to 91.5% (Malinsky et al., 2018) of the total genome. For instance, some studies used GATK's CallableLoci tool (Van der Auwera et al., 2013) that estimates the number of sites that pass the DP filters from the read alignment (.bam) files (e.g., Wu et al., 2020) while another study (Wang et al., 2020) used the variant calling files (.vcf) from the samtools mpileup caller (Li et al., 2009). From GATK 4 onward, CallableLoci is no longer supported, yet, with the BP_RESOLUTION mode, every single site of the genome has a depth and genotype quality value that can be used to estimate the callable sites (used in e.g., Bergeron et al., 2021; Pfeifer, 2017). Moreover, some studies restrict the CG to the orthologous genome in order to match for base composition when making comparisons across species (e.g., Wu et al., 2020). Due to these differences, it is important to report which methodology and filters are used to estimate CG.

Assessing false-negatives. On the number of sites considered callable, additional corrections for the FNR can be included. Indeed, even if the CG represents the sites that pass most of the individual filters, some filters can simply not be applied to non-polymorphic sites. The

519 methods and results differ between studies, with an estimated FNR from 0 (Smeds et al.,
520 2016; Tatsumoto et al., 2017) to 44% (Thomas et al., 2018).

521 One way to estimate a FNR is to introduce random DNMs to the sequencing reads and run
522 the entire pipeline (steps 2-4) to calculate its efficiency in finding these simulated DNMs
523 (e.g., Feng et al., 2017; Jónsson et al., 2017; Pfeifer, 2017; Wu et al., 2020). The false-
524 negative rate can then be estimated as:

$$FNR = \frac{de\ novo_{missed}}{de\ novo_{simulated}}$$

525 This method corrects for errors during alignment, post-alignment processing, calling, and
526 filtering as the reads are passed into the pipeline a second time. However, it can be
527 computationally intensive as variant calling needs to be run multiple times and is a resource
528 and time-intensive step.

529 Another way to estimate the FNR is to use the number of callable sites that will be filtered
530 away by filters different from those taken into account in the CG estimation, such as site or
531 allelic balance filters (Bergeron et al., 2021; Besenbacher et al., 2019; Thomas et al., 2018).
532 As some site filters are inferred during variant calling based on statistical tests following
533 known null distributions, it is possible to estimate the proportion of callable sites filtered
534 away by these site filters (Bergeron et al., 2021; Besenbacher et al., 2015). Moreover, some
535 true DNMs could have an allelic balance outside the allelic balance filter chosen due to
536 sequencing variability or mosaicism. This bias can be estimated by the heterozygous sites in
537 the offspring (that are not DNMs) presenting an allelic balance outside the allelic balance
538 filter, assuming that this bias occurs at the same rate at DNMs and heterozygous sites in the
539 offspring (i.e. one parent is homozygous for the reference allele, one parent is homozygous
540 for the alternative allele, and the offspring heterozygous). Therefore, FNR can be inferred as
541 the proportion of true heterozygous sites outside the AB filter as:

542 $$FNR = \frac{True\ heterozygous\ sites_{outside\ AB}}{True\ heterozygous\ sites}.$$

543 Finally, the denominator can be estimated based on a probability to detect a DNM at a site,
544 given various parameters at that site. Thus, there is no clear distinction between *CG* and FNR,
545 as the latter is part of the *CG* estimation. Specifically, Besenbacher et al. (2019) used
546 inherited variants to estimate the probability that a DNM at a given site would pass all filters

conditional on the depth of each individual. They then summed these probabilities to calculate the number of callable sites in the genome.

Mutationathon: two-fold variation in estimated rates from the same trio

To understand the effect of various methods on mutation rate estimates from a single dataset, a three-generation pedigree of rhesus macaque (*Macaca mulatta*) was analyzed by researchers from five groups: Lucie Bergeron (LB), Søren Besenbacher (SB), Cyril Versoza (CV), Tychele Turner (TT), and Richard Wang (RW). The macaque pedigree consisted of Noot (father), M (mother), Heineken (daughter), and Hoegaarde (Heineken's daughter) (Figure 3a). Each individual was sequenced with BGISEQ-500 at an average coverage between 40X (Noot) and 70X (all other individuals). The raw data were trimmed using SOAPnuke (Chen et. al, 2017) to remove adaptors, low-quality reads, and N-reads (see Material and Methods for more information). Trimmed reads were shared with all participants, who applied their respective pipelines to identify DNMs in Heineken and to estimate a per-site per-generation germline mutation rate.

Each group of investigators implemented their own set of filters (Supplementary Table 4) and detected between 18 (CV) and 32 (SB) candidate DNMs. After PCR amplification and Sanger sequencing validation of the DNM candidates from all research groups (43 distinct sites), we validated 33 positions as true positive DNMs, six were determined to be false-positive calls, and four did not successfully amplify (Figure 3b and Supplementary Table 5). No group found all true positive DNMs. Of the 33 true positive DNMs, only 7 were detected by all research groups (Figure 3c). Fourteen additional true positive mutations were detected by at least four groups; 6 detected by all except CV, 4 by all except RW, 2 by all except LB, 1 by all except SB, and 1 by all except TT. Of the 12 remaining true positive mutations, 5 were detected by three groups, 1 by two groups, and 6 by a single group. The candidate DNMs found by a single group are more likely to be false positives as the six false-positive candidates revealed by the PCR experiment were all candidates detected by a single pipeline. The transmission rate to the next generation varied between 52% (with SB pipeline: 15 true positive DNMs transmitted on 29 true positive candidates) and 67% (with RW pipeline: 14 true positive DNMs transmitted on 21 true positive candidates). The transmission rate of all true positive DNMs (33) was 67% with 21 DNMs transmitted to the next generation; this rate

is not significantly different from the expected 50% inheritance (binomial test p-value = 0.08).

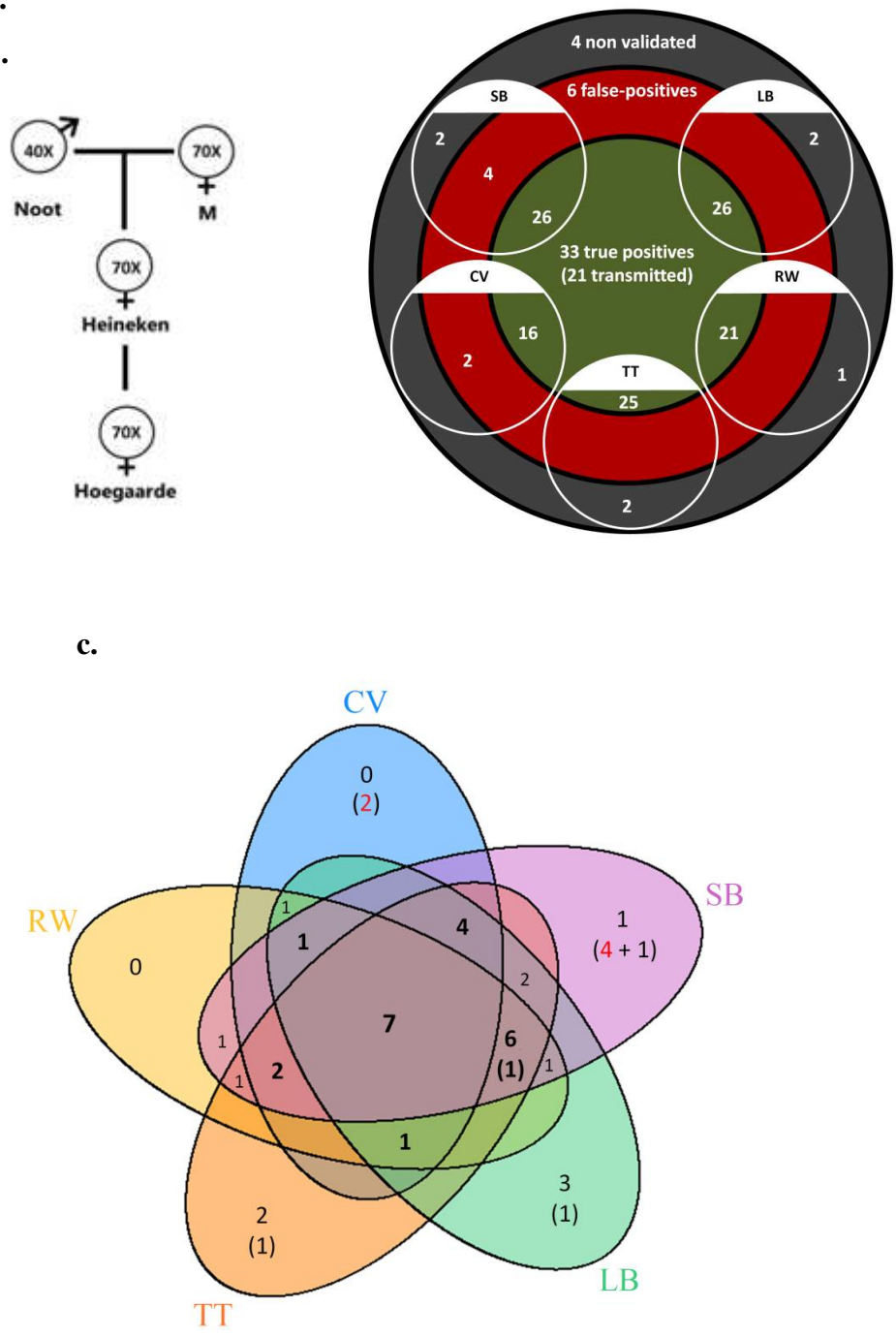

Figure 3 – Candidate DNMs from the Mutationathon. a. The pedigree of three generations of rhesus macaques was sequenced and shared with five groups of researchers. Sequencing coverage is indicated for each individual. b. The five groups (LB: Lucie Bergeron, SB: Søren Besenbacher, CV: Cyril Versoza, TT:

Tychele Turner, RW: Richard Wang) detected a total of 43 candidate DNMs in Heineken. The PCR amplification and Sanger sequencing validation showed that 33 of those candidates were true positive DNMs, six were false-positive calls, and four did not successfully amplify. c. Venn diagram of the mutations found by each research group. In bold are the candidates shared by at least four different groups. Between brackets are the candidates that were not validated by the PCR experiment either because they did not successfully amplify (in black) or because the sequencing revealed false-positive calls (in red). See Material and Methods for details on the experiment and Supplementary Figure 1 for the results of the PCR experiment.

In addition to identifying DNMs, each group was tasked with estimating the per-site per generation rate of mutation. The final estimated rate depends on the size of the callable genome (CG) considered by each group, as well as corrections for false positives and false negatives. Even with the variation in the number of candidate DNMs from each group (Figure 4a), different values for these additional parameters could still have resulted in equivalent rate estimates between different groups. However, differences in methodology led to almost a two-fold variation in the estimated rates, greater than the variation in the number of DNMs. TT estimated the lowest rate of 0.46×10^{-8} mutations per-site per generation (Figure 4b). This estimate was based on autosomes and the X chromosome (where two candidates were found), and the CG represented almost the full genome size. Using the full genome size in the denominator is commonly used in human studies, for which most of the genome is callable due to the high-quality reference genome, while stricter corrections are usually applied in non-human studies. CV, RW, SB, and LB found similar rates, with 25% differences between the lowest and the highest rate and large overlap of the confidence intervals (Figure 4c). RW estimated the highest rate with 0.85×10^{-8} mutations per-site per generation, from a relatively small set of candidates (22), yet the denominator was also small as CG represented about 50% of the autosomal genome. SB and LB estimated a similar value of CG, representing approximately 80% of the autosomal genome; however, there was a difference in rates due to the smaller number of candidates found by LB (28) compared to SB (32).

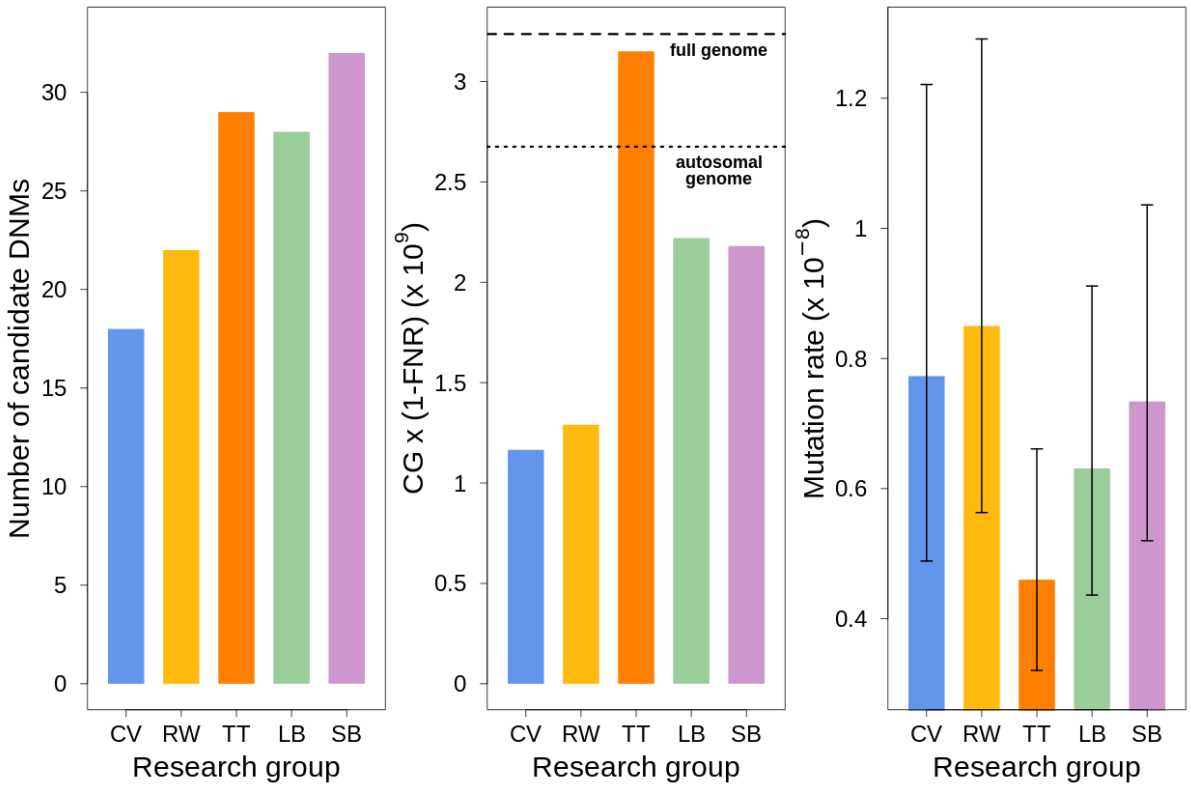

Figure 4 – Estimated germline mutation rates from the Mutationathon. a. Number of candidates DNMs found by each group (TT found 2 candidates on a sex chromosome). b. Estimation of the denominator (i.e. the callable genome corrected by the FNR) by each group. c. Estimated mutation rate per-site per generation, the error bars correspond to the confidence intervals for binomial probabilities (calculated using the R package 'binconf').

The different individual filters applied by each group explain some of the differences in the candidate DNMs (Table 2 and Supplementary Table 4). For instance, many groups filtered away candidate sites where the parents were heterozygotes, as they could be more prone to false-positive calls. TT's pipeline was the only one to find a candidate mutation at a site where the father was heterozygous C/G, the mother was homozygous for the reference allele G/G, and the offspring was heterozygous A/G. These genotypes were validated by the PCR

experiment, indicating that a true germline mutation has arisen at a heterozygous site in a parental genome. Each method varied in power to detect the true DNMs (sensitivity), and in the proportion of validated true candidates on the overall candidates found (precision). For instance, RW used especially conservative filters on the allelic balance for both the offspring (AB) and the number of alternative alleles allowed in the parents (AD). It resulted in a lower sensitivity, only 22 candidates were found, but a high precision as no candidates were determined to be false-positive calls. Similarly to RW, some groups were conservative on the AB filter, while other groups were more conservative on the GQ filter (SB and LB) or DP filter (LB, CV, RW). For instance, SB used a relaxed filter on DP, with a minimum threshold of 10X, but a relatively conservative threshold on AB and GQ criteria. TT did not use strict filters for any parameter, however, the precision was increased by the required overlap among multiple variant callers.

Table 2 – Individual filters used by the different groups to detect DNMs in Heineken (difference in the other steps of the pipeline in Supplementary Table 4)**.**

Research group	**Candidates DNMs**	**DP filter**	**GQ filter**	**AD filter**	**AB filter**	**Additional filters**
CV	18	$0.5 \times dp_{ind}$ - $2 \times dp_{ind}$	40	0	0.25 - 0.75	
RW	22	20 - 80	20	0	0.35	Alternative allele on both strands
TT	27	Minimum 10 X	20	0	0.25	Overlap 3 different variant callers
LB	28	$0.5 \times dp_{ind}$ - $2 \times dp_{ind}$	60	none	0.3 - 0.7	Manual curation (6 candidates removed)

SB	32	Minimum 10 X – Maximum 1.75 × dp_{ind}	55	0	0.3	Alternative allele in both strands. lowQ AD2 > 1

667

668 We explored the effect of the individual filter on the number of candidate DNMs, the number
669 of false-positive calls (FP), the callable genome (CG), the false-negative rate (FNR), and the
670 final estimated mutation rate per-site per generation (μ). We used the LB pipeline (see
671 individual filters in Table 2 and other methods in Supplementary Table 4) and changed one
672 filter at a time using various criteria used by the Mutationathon participants and in the
673 literature (Figure 5 and Supplementary Table 6). The GQ filter had the largest impact on the
674 number of mutations and the final estimated mutation rate. The number of candidate DNMs
675 found with GQ < 20 was three times higher than the one obtained with the most conservative
676 GQ filter (GQ_{Hom} < 100 and GQ_{Het} < 200), and the difference was still two-fold after
677 correcting for FP calls. The callable genome (CG) also decreased with GQ < 80, leading to an
678 estimated rate 39% lower when GQ < 80 ($\mu = 0.56 \times 10^{-8}$) compared to when GQ < 20 ($\mu =$
679 0.91×10^{-8}). This filter also seems to be the most efficient at reducing the number of FP calls,
680 estimated here with the manual curation method, as more than 90% of the candidates DNMs
681 were false positives with no GQ filter while we found no false positives with conservative
682 GQ filters (GQ < 80 and GQ_{Hom} < 100 and GQ_{Het} < 200). Another important filter was the
683 allelic balance on the heterozygous offspring, resulting in a two-fold difference in the number
684 of candidate DNMs detected, and 1.5-fold difference after the correction for FP calls. Yet, the
685 estimated FNR was almost five times higher when using a conservative AB filter (AB < 0.4
686 and AB > 0.6; FNR = 15.8%) compared to the least conservative AB filter (AB < 0.2; FNR =
687 3.5%). This led to a mutation rate estimate 28% lower with the conservative AB filter (AB <
688 0.4 and AB > 0.6; $\mu = 0.69 \times 10^{-8}$ and AB < 0.2; $\mu = 0.83 \times 10^{-8}$). The DP filter also impacted
689 the estimated rate but to a lesser extent with only a 6% difference between the estimated
690 mutation rate with DP < 10 ($\mu = 0.67 \times 10^{-8}$) and the most conservative DP filter (DP < 0.5 ×
691 $depth_{individual}$ and DP > 2 × $depth_{individual}$; $\mu = 0.63 \times 10^{-8}$). Finally, the AD filter did not show
692 a large impact on the mutation rate, with less than 2% difference between no filter on AD (μ
693 $= 0.63 \times 10^{-8}$) and the conservative AD > 0 ($\mu = 0.62 \times 10^{-8}$).

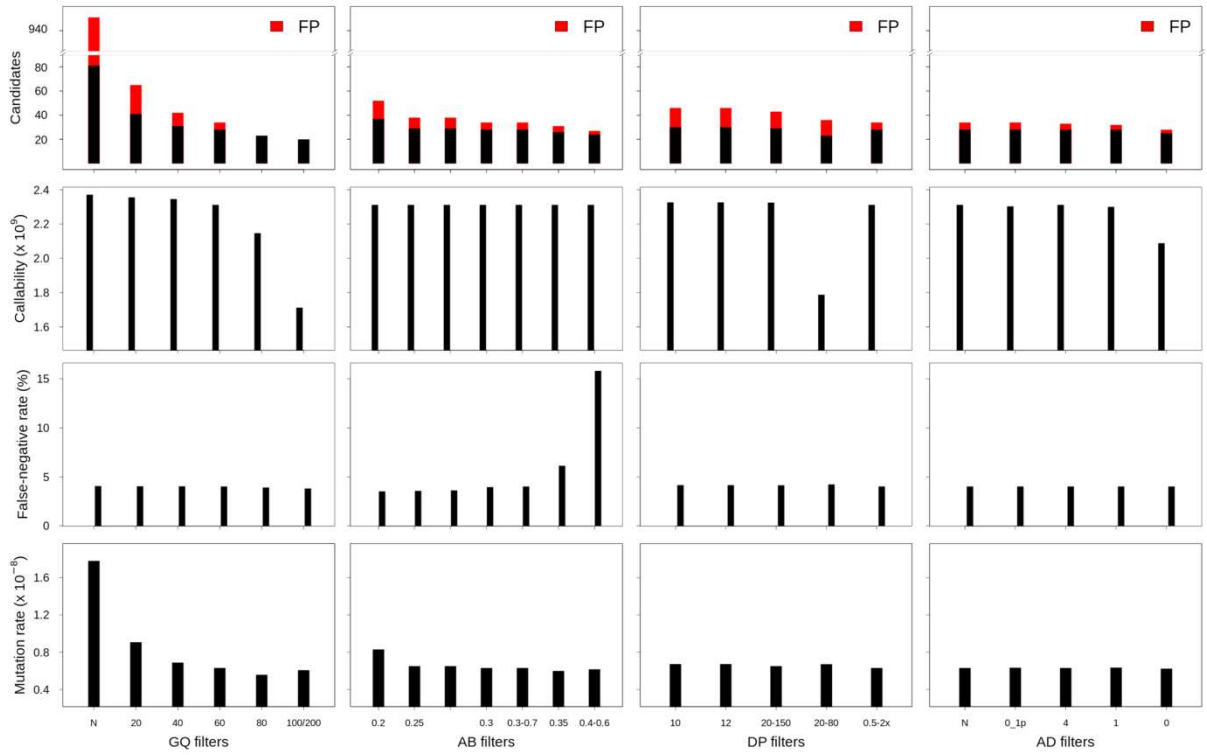

Figure 5 – The impact of individual filters on the estimated rate of a trio of rhesus macaques. The default filters used by LB pipeline were: DP < 0.5 × depth $_{individual}$; DP > 2 × depth$_{individual}$; GQ < 60; AB < 0.3; AB > 0.7, no AD filter. The mutation rate was calculated with LB pipeline as $\mu = \frac{nb_{candidate\ DNMs} - FP}{2 \times CG \times (1-FNR)}$.

These results show that some of the differences in estimated rates between the five research groups may be attributed to the individual filters. Yet, earlier steps in the different bioinformatic pipelines could also lead to differences in candidate DNMs and estimated rates. For instance, the site filters were different between some of the groups (see Supplementary Table 4). Testing different combinations of site filters on the shared trio of rhesus macaques affected the set of SNPs detected, which could lead to variation in candidate DNMs detected. For instance, on the 12,634,956 variants found by LB pipeline, 473,142 SNPs were removed when using GATK advised filters (QD < 2.0; MQ < 40.0; FS > 60.0; SOR > 3.0; MQRankSum < -12.5; ReadPosRankSum < -8.0), while the stringent filters used by LB pipeline (QD < 2.0, FS > 20.0, MQ < 40.0, MQRankSum < - 2.0, MQRankSum > 4.0, ReadPosRankSum < - 3.0, ReadPosRankSum > 3.0 and SOR > 3.0) removed 1,124,005 SNPs. Despite this difference in the number of SNPs, using the LB pipeline to detect candidate DNMs on the three callset (no filters, GATK advised filter or stringent filters), led

to the same final number of candidate DNMs due to the stringent individual filters applied in the following steps of the pipeline. Other steps, such as mapping and variant calling, could also lead to some of the differences between the five groups. For instance, the six candidates identified as false positives by the Sanger sequencing were filtered away in the LB pipeline. Four of the false-positive candidates were not detected because all individuals were genotyped as homozygous for the reference allele, one position was filtered out by the mapping quality site filter (MQ < 40), one position had DP = 0. Thus, differences in the mapping of the reads and variant callers explain some of the discrepancies between pipelines.

Overall, these results show that for the same dataset, differences in estimated mutation rates caused by methodological discrepancies are non-negligible. Therefore, such differences should be considered when comparing mutation rates between different species when they are estimated by different pipelines. Some of the differences in estimated rates between groups can be attributed to the different individual filters applied for the detection of candidate DNMs. Most notably, varying the GQ and AB filters leads to large variations in estimated rates. Some of the difference is also introduced in earlier steps when mapping reads and calling variants. Moreover, the estimated callable genome is different between the five groups; in addition to changing the denominator of the mutation rate calculation, this difference could reflect the ability of individual methods to query mutation in different genomic regions. Some variation might therefore be explained by true mutation rate heterogeneity between genomic regions (such as low or high complex regions). Our results also suggest that despite the different methods and filters, the estimated rates are comparable when both the numerator (number of candidates and false positives) and the denominator (CG and false-negative rate) are carefully corrected. For instance, CV, SB, LB, and RW estimated similar rates, but SB and RW used a probabilistic method to calculate the CG, while LB used strict filters (DP and GQ) on a base-pair resolution variant calling file (vcf) and corrected for FNR using the site filters and the allelic balance filter and CV used a similar method to estimate CG, yet, did not apply a correction for FNR.

Best practices

When estimating germline mutation rates from pedigree samples, there is no standardized set of methods. Different studies use different software versions and filtering thresholds, which

744 can impact the estimated rate and can complicate the comparison of rates between or within
745 species across studies (in addition to the biological variation introduced by the age of the
746 parents used in each study; Table 1). Here, we provide guidelines for each step in DNM
747 calling and rate estimation. However, we note that sample quality, reference genomes, and
748 other technical factors differ across studies and thus require study- or species-specific
749 thresholds. Therefore, it is advised to report the methodology used in a standardized way.
750 Table 3 proposes a checklist of parameters that should be reported in studies of germline
751 mutation rates.

752

753 **Table 3 – Information that should ideally be reported when presenting**
754 **results on DNMs**. See Supplementary Table 4 for an example of this table filled
755 out for the five pipelines used to analyze the trio of rhesus macaques.

Step of the analysis	**Information to report**
1 – Sampling and sequencing	Type of sample (tissue, etc.)
	Storage duration, buffer, temperature
	Type of library preparation
	Average sequencing coverage
	Sequencing technology and read lengths
2 – Alignment and post-alignment processing	Trimming of adaptors and low-quality reads
	Reference assembly version Autosomes only or whole genome?
	Mapping software and version
	Duplicate removal software and version
	Base quality score recalibration (yes/no)
	If yes, which type of data used as known variants

	Realignments around indels?
	Other filters?
3 – Variant calling	Software and version
	Mode: joint genotyping? Gvcf blocks? Gvcf in base-pair resolution?
4 – Detecting ***de novo*** **mutations**	Site filters on vcf files and justification
	Individual filters, threshold, and remaining candidates after each filter
	False discovery rate estimation method: PCR validation? Manual curation? Transmission rate deviation?
5 – Mutation rate estimation	Callable genome estimation method: File used? Filters taken into account?
	False-negative rate estimation method: simulation? Filters? Probability?

Moreover, some benchmarks could be helpful to ease the comparison between studies such as:

- the transition-to-transversion ratio (ti/tv),
- the spectrum of mutations (see Supplementary Figure 2 for an example from the Mutationathon),
- percentage of mutations in CpG locations,
- base composition (percentage of A/T or C/G),
- nucleotide heterozygosity in unrelated individuals,
- if population data are available, the number of DNMs that are in known SNPs of the population,

- the contribution of each sex to the total number of mutation bias when phasing of mutations is possible,
- transmission rate to the next generation when extended trios are available,
- the average age of the parents at the time of reproduction, if known
- distribution of the allelic balance of true heterozygotes, candidate DNMs after all filters except the allelic balance, and the final set of candidate DNMs.

Conclusion and perspectives

Different filters can lead to differences in estimated rates, which emphasizes the difficulty in comparing pedigree-based germline mutation rates estimated from different studies. The variation observed could be partially due to the biology and life-history traits of species, but some of the variations will also be caused by methodological differences. Here, we provided some best practices that can be used when estimating germline mutation rates from pedigree samples. However, it is hard to provide hard cutoffs of filters that apply to every situation, and we advise choosing appropriate filters depending on the data available. We have also raised some points that should be addressed in individual studies, such as estimation of the false discovery rate, false-negative rate, and the callable genome size. Nevertheless, more exploration should be done to understand the best strategy for the different steps required in every study of the mutation rates. Without a clear consensus on approaches for estimating the germline mutation rate, it seems that the best strategy will be to carefully report all methods and parameters used. The trio of rhesus macaque used in this analysis is publically available, along with the validated candidate DNMs, and could serve as a resource for testing new strategies. On a more positive note, it is important to point out that two recent, independent studies of the per-generation mutation rate in rhesus macaque reported rates that were within 5% of each other for individuals of the same age (Bergeron et al., 2021; Wang et al., 2020). We hope that careful studies using a variety of methods will be able to similarly arrive at accurate estimates of important biological parameters.

With the growing number of studies on pedigree-based estimation of germline mutation rate, some directions that have been neglected could be explored. For instance, even when the sample size is large, most studies use samples originating from small geographic regions; it would be of great interest to further explore potential variation in mutation rates across

diverse populations (e.g. Kessler et al., 2020). Most studies are conducted on genomic DNA collected from somatic tissues. As a result, if samples come from only a single trio, one cannot distinguish early postzygotic mutations occurring in the offspring from germline mutations in the parents. While mutations occurring early enough in offspring development will be passed on to the next generation--and should therefore still be considered *de novo* mutations--they will behave differently from mutations arising in the parental generation. For instance, we will not expect an increase of these mutations with parental age (Jónsson et al., 2018). Therefore, it is of interest to distinguish between these two types of mutation, especially for biomedical research. A possible way to discard those mutations would be to compare somatic and germline cells from the same individual. However, extracting DNA directly from sperm and eggs can be challenging, especially for non-human species, limiting the application of this strategy. Another area for additional future work is to look at *de novo* structural variants. As they are even rarer than SNPs, it is hard to detect them over a single generation. Yet, with the growing number of trios and generations considered in recent studies, it would be of interest to quantify and describe those DNMs as well (e.g. Belyeu et al., 2021; Thomas et al., 2021). The development of accurate long-read sequencing technologies also offers opportunities for better detection of DNMs and *de novo* structural variants. Finally, most studies on non-human species only explore the autosomal chromosomes, largely because important filters such as allelic balance cannot be used on the sex chromosomes in both sexes. However, given the consistent differences observed between species in the rate of evolution on autosomes and sex chromosomes (e.g. Wilson Sayres and Makova, 2011), it would be very interesting to look more closely at the per-generation mutation rate on sex chromosomes.

Material and methods

Mutationathon sequences. The pedigree used for the Mutationathon was previously sequenced as part of a larger project on the mutation rate of rhesus macaques (BioProject: PRJNA588178; Bergeron et al., 2021). Nine lanes were used in this analysis (three lanes for the father and two lanes for the other individuals) and are publically available on NCBI:

- CL100066413_L01 (SRA run SRR10426295), mother M
- CL100089164_L01 (SRA run SRR10426294), mother M
- CL100078308_L01 (SRA run SRR10426275), father Noot

- CL100078335_L01 (SRA run SRR10426264), father Noot
- CL100078335_L02 (SRA run SRR10426253), father Noot
- CL100066412_L02 (SRA run SRR10426291), offspring Heineken
- CL100095002_L02 (SRA run SRR10426290), offspring Heineken
- CL100066408_L01 (SRA run SRR10426256), next generation offspring Hoegaarde
- CL100094917_L01 (SRA run SRR10426255), next generation offspring Hoegaarde

A trimming step was done on all sequences to remove the adaptors (allowing a mismatch of two bases), the low-quality reads (with more than 5% of N bases or a base quality score < 10 in more than 20% of the read), and the reads smaller than 60 bases after the quality control. Trimming was done using SOAPnuke version 1.5.6 (Chen et al., 2017), with the following command:

> *SOAPnuke filter -f AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA -r AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG -1* sequence_read_1 *-2* sequence_read_2 *-G -Q 2 -l 10 -q 0.2 -E 60 -5 0 -M 2 -o* sequence_clean *-C* sequence_read_1_clean *-D* sequence_read_2_clean

Each group implemented its pipeline to estimate a rate (details are provided in Supplementary Table 4).

Data analysis. The comparison of each individual filter was done using LB pipeline, changing one filter at a time and recalculating the number of candidates DNMs detected, the potential false-positive candidates with the manual curation method, the callable genome, the FNR on the allelic balance filter and site filters, and the mutation rate per site per generation. The comparison of the site filters was also done on the SNPs found by LB pipeline.

PCR experiment and Sanger resequencing. We designed multiple sets of primers for the 43 candidate sites on NCBI primer blast tool (Ye et al., 2012: https://www.ncbi.nlm.nih.gov/tools/primer-blast/). In some cases, sequencing primers were adjusted to avoid sequencing failure due to poly-AAA or TTT runs. PCRs were carried out in 25µL volumes [2.5 units Dream Taq DNA Polymerase (Thermo Scientific), 1X Dream Taq Green Buffer, 0.2mM dNTPs, 2–3mM MgCl2, 2.5- 44 ng DNA template, filled to 25µL with double-distilled (ddH2O) water]. Thermocycling was performed in a BIORAD PTC-100 thermocycler. The cycle program comprised of an initial denaturation at 95°C for 2min, followed by 35 cycles of 15sec at 95°C, 15sec at 52°C- 55°C, and 30sec at 72°C. Cycling

was terminated with a 5min extension at 72°C. PCR products were purified using commercially available spin columns (Invitek) or PureIT ExoZap PCR Clean-up (Ampliqon). Sanger sequencing was conducted at Eurofins Genomics, Europe using the primers of the amplification procedure using both forward and reverse primers. In Supplementary Figure 1, the chromatograms with the best base quality value are provided. Supplementary Table 7 provides details about the primers and accession number of the sequences on GenBank.

Data and code availability

All the sequences used for the Mutationathon were previously generated and released in NCBI (Bergeron et al., 2021). The sequences used were for the mother M (BioSample SAMN13230631), lanes CL100066413_L01 (SRA run SRR10426295) and CL100089164_L01 (SRA run SRR10426294); for the father Noot (BioSample SAMN13230623): lanes CL100078308_L01 (SRA run SRR10426275), CL100078335_L01 (SRA run SRR10426264) and CL100078335_L02 (SRA run SRR10426253); for the offspring Heineken (BioSample SAMN13230633): lanes CL100066412_L02 (SRA run SRR10426291) and CL100095002_L02 (SRA run SRR10426290); and for the second generation offspring Hoegaarde (BioSample SAMN13230649): lanes CL100066408_L01 (SRA run SRR10426256) and CL100094917_L01 (SRA run SRR10426255). The Sanger sequences generated during the PCR validation, were deposited on GenBank under the accession number MZ661796 - MZ662076.

The scripts used by the participants of the Mutationathon are publically available:

- CV: https://github.com/PfeiferLab/mutationathon;
- RW: https://github.com/Wang-RJ/mutationathon;
- TT: Wilfert, A. B., Turner, T. N., Murali, S. C., Hsieh, P., Sulovari, A., Wang, T., ... & Eichler, E. E. 2021. Recent ultra-rare inherited variants implicate new autism candidate risk genes. *Nature Genetics*, 53(8):1125-1134. doi: 10.1038/s41588-021-00899-8;
- LB: https://github.com/lucieabergeron/germline_mutation_rate;
- SB: https://github.com/besenbacher/GreatApeMutationRate2018

Acknowledgments

We would like to thank GenomeDK at Aarhus University and Arizona State University's Research Computing for providing computational resources and support for the LB pipeline and CV pipeline respectively. We also thank Hákon Jónsson for helpful comments on the manuscript and Maria Kamilari for helpful input on the PCR validation experiment. SPP is supported by a US National Science Foundation CAREER grant (DEB-2045343).

Competing Interests

The authors declare that they have no competing interests.

References

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. 2005. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* **71**:8966–8969. doi:10.1128/AEM.71.12.8966-8969.2005

Acuna-Hidalgo R, Veltman JA, Hoischen A. 2016. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol*. doi:10.1186/s13059-016-1110-1

Auwera G Van der, O'Connor B. 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra.

Baust JG. 2008. Strategies for the Storage of DNA. *Biopreserv Biobank* **6**:251–252. doi:10.1089/BIO.2008.0604.LETT

Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, Gupta M, Nicholas TJ, Brown J, Baird L, Devlin B, Sanders SJ, Jorde LB, Talkowski ME, Quinlan AR. 2021. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am J Hum Genet* **108**:597–607. doi:10.1016/J.AJHG.2021.02.012

Bergeron LA, Besenbacher S, Bakker J, Zheng J, Li P, Pacheco G, Sinding M-HS, Kamilari M, Gilbert MTP, Schierup MH, Zhang G. 2021. The germline mutational process in rhesus macaque and its implications for phylogenetic dating. *Gigascience* **10**:1–14.

doi:10.1093/GIGASCIENCE/GIAB029

Besenbacher S, Hvilsom C, Marques-Bonet T, Mailund T, Schierup MH. 2019. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat Ecol Evol* **3**:286–292. doi:10.1038/s41559-018-0778-x

Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R, Rubio-García A, Lescai F, Demontis D, Rao J, Ye W, Mailund T, Friborg RM, Pedersen CNS, Xu R, Sun J, Liu H, Wang O, Cheng X, Flores D, Rydza E, Rapacki K, Damm Sørensen J, Chmura P, Westergaard D, Dworzynski P, Sørensen TIA, Lund O, Hansen T, Xu X, Li N, Bolund L, Pedersen O, Eiberg H, Krogh A, Børglum AD, Brunak S, Kristiansen K, Schierup MH, Wang J, Gupta R, Villesen P, Rasmussen S. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6**:5969. doi:10.1038/ncomms6969

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, Gudjonsson SA, Magnusdottir DN, Jonasdottir Aslaug, Jonasdottir Adalbjorg, Kristjansson RP, Sverrisson ST, Holley G, Palsson G, Stefansson OA, Eyjolfsson G, Olafsson I, Sigurdardottir O, Torfason B, Masson G, Helgason A, Thorsteinsdottir U, Holm H, Gudbjartsson DF, Sulem P, Magnusson OT, Halldorsson B V, Stefansson K. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**:779–786. doi:10.1038/s41588-021-00865-4

Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, Barrera DJ, Lin GN, Malhotra D, Watts AC, Wong LC, Estabillo JA, Gadomski TE, Hong O, Fajardo KVF, Bhandari A, Owen R, Baughn M, Yuan J, Solomon T, Moyzis AG, Maile MS, Sanders SJ, Reiner GE, Vaux KK, Strom CM, Zhang K, Muotri AR, Akshoomoff N, Leal SM, Pierce K, Courchesne E, Iakoucheva LM, Corsello C, Sebat J. 2016. Frequency and Complexity of De Novo Structural Mutation in Autism. *Am J Hum Genet* **98**:667–679. doi:10.1016/J.AJHG.2016.02.018

Campbell CR, Tiley GP, Poelstra JW, Hunnicutt KE, Larsen PA, Lee H-J, Thorne JL, dos Reis M, Yoder AD. 2021. Pedigree-based and phylogenetic methods support surprising patterns of mutation rate and spectrum in the gray mouse lemur. *Heredity (Edinb)* **127**:233–244. doi:10.1038/s41437-021-00446-5

Chen J, Li X, Zhong H, Meng Y, Du H. 2019. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci Rep* **9**:1–13. doi:10.1038/s41598-019-45835-3

950 Chen Yuxin, Chen Yongsheng, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z, Zhang X,
951 Wang J, Yang H, Fang L, Chen Q. 2017. SOAPnuke: A MapReduce acceleration-supported
952 software for integrated quality control and preprocessing of high-throughput sequencing data.
953 *Gigascience* **7**:1–6. doi:10.1093/gigascience/gix120

954 Conrad DF, Keebler JEM, Depristo MA, Lindsay SJ, Zhang Y, Cassals F, Idaghdour Y, Hartl CL,
955 Torroja C, Garimella K V, Zilversmit M, Cartwright R, Rouleau G, Daly M, Stone EA, Hurles
956 ME, Awadalla P. 2011. Variation in genome-wide mutation rates within and between human
957 families. *Nat Genet* **43**:712. doi:10.1038/ng.862

958 Eggertsson HP, Jónsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, Zink F, Hjorleifsson
959 KE, Jonasdottir Aslaug, Jonasdottir Adalbjorg, Jonsdottir I, Gudbjartsson DF, Melsted P,
960 Stefansson K, Halldorsson B V. 2017. Graphtyper enables population-scale genotyping using
961 pangenome graphs. *Nat Genet* **49**:1654–1660. doi:10.1038/ng.3964

962 Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT,
963 Gudbjartsson DF, Stefansson K, Halldorsson B V, Melsted P. 2019. GraphTyper2 enables
964 population-scale genotyping of structural variation using pangenome graphs. *Nat Commun* **10**:1–
965 8. doi:10.1038/S41467-019-13341-9

966 Feng C, Pettersson M, Lamichhaney S, Rubin C-J, Rafati N, Casini M, Folkvord A, Andersson L.
967 2017. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation
968 rate. *Elife* **6**:e23907. doi:10.7554/eLife.23907.001

969 Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I. 2015. Genome-wide patterns and
970 properties of de novo mutations in humans. *Nat Genet* **47**:822. doi:10.1038/ng.3292

971 Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R.
972 2013. Quantifying population genetic differentiation from next-generation sequencing data.
973 *Genetics* **195**:979–992. doi:10.1534/genetics.113.154740

974 Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv*
975 *Prepr arXiv* **1207.3907**.

976 GATK team. 2021. Base Quality Score Recalibration (BQSR). *Genome Anal Toolkit Doc*.
977 https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-
978 Recalibration-BQSR-

979 GATK team. 2020. I am unable to use VQSR (recalibration) to filter variants. *Genome Anal Toolkit*
980 *Doc*. https://gatk.broadinstitute.org/hc/en-us/articles/360037499012-I-am-unable-to-use-VQSR-

recalibration-to-filter-variants

Halldorsson B V., Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, Gudjonsson SA, Frigge ML, Thorleifsson G, Sigurdsson A, Stacey SN, Sulem P, Masson G, Helgason A, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**:eaau1043. doi:10.1126/science.aau1043

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, Mullaart E, Coppieters W, Georges M. 2017. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. *bioRxiv* 079863. doi:10.1101/079863

Jónsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, Zink F, Helgason EA, Jonsdottir I, Gylfason A, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Beyter D, Steingrimsdottir T, Norddahl GL, Magnusson OT, Masson G, Halldorsson B V., Thorsteinsdottir U, Helgason A, Sulem P, Gudbjartsson DF, Stefansson K. 2021. Differences between germline genomes of monozygotic twins. *Nat Genet* **53**:27–34. doi:10.1038/s41588-020-00755-1

Jónsson H, Sulem P, Arnadottir GA, Pálsson G, Eggertsson HP, Kristmundsdottir S, Zink F, Kehr B, Hjorleifsson KE, Jensson BÖ, Jonsdottir I, Marelsson SE, Gudjonsson SA, Gylfason A, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Stacey SN, Magnusson OT, Thorsteinsdottir U, Masson G, Kong A, Halldorsson B V., Helgason A, Gudbjartsson DF, Stefansson K. 2018. Multiple transmissions of de novo mutations in families. *Nat Genet* **50**:1674. doi:10.1038/s41588-018-0259-9

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H, Gylfason A, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Rafnar T, Frigge M, Stacey SN, Th. Magnusson O, Thorsteinsdottir U, Masson G, Kong A, Halldorsson B V., Helgason A, Gudbjartsson DF, Stefansson K. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**:519–522. doi:10.1038/nature24018

Karczewski KJ, Gauthier LD, Daly MJ. 2019. Technical artifact drives apparent deviation from Hardy-Weinberg equilibrium at CCR5-Δ32 and other variants in gnomAD. *bioRxiv* 784157. doi:10.1101/784157

Kessler MD, Loesch DP, Perry JA, Heard-Costa NL, Taliun D, Cade BE, Wang H, Daya M, Ziniti J, Datta S, Celedón JC, Soto-Quiros ME, Avila L, Weiss ST, Barnes K, Redline SS, Vasan RS, Johnson AD, Mathias RA, Hernandez R, Wilson JG, Nickerson DA, Abecasis G, Browning SR,

1013 Zöllner S, O'Connell JR, Mitchell BD, O'Connor TD. 2020. De novo mutations across 1,465
1014 diverse genomes reveal mutational insights and reductions in the Amish founder population.
1015 *Proc Natl Acad Sci U S A* **117**:2560–2569. doi:10.1073/pnas.1902766117

1016 Koch EM, Schweizer RM, Schweizer TM, Stahler DR, Smith DW, Wayne RK, Novembre J. 2019.
1017 De Novo Mutation Rate Estimation in Wolves of Known Pedigree. *Mol Biol Evol* **36**:2536–
1018 2547. doi:10.1093/molbev/msz159

1019 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson
1020 A, Jonasdottir Aslaug, Jonasdottir Adalbjorg, W Wong WS, Sigurdsson G, Bragi Walters G,
1021 Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Th Magnusson O,
1022 Thorsteinsdottir U, Stefansson K. 2012. Rate of de novo mutations and the importance of
1023 father's age to disease risk. *Nature* **488**:471. doi:10.1038/nature11396

1024 Lescai F, Marasco E, Bacchelli C, Stanier P, Mantovani V, Beales P. 2014. Identification and
1025 validation of loss of function variants in clinical contexts. *Mol Genet Genomic Med* **2**:58–63.
1026 doi:10.1002/mgg3.42

1027 Li H. 2020. Base quality scores are essential to short read variant calling.
1028 *http://lh3.github.io/2020/05/27/base-quality-scores-are-essential-to-short-read-variant-calling*.
1029 http://lh3.github.io/2020/05/27/base-quality-scores-are-essential-to-short-read-variant-calling

1030 Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples.
1031 *Bioinformatics*. doi:10.1093/bioinformatics/btu356

1032 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform.
1033 *Bioinformatics* **25**:1754–176010. doi:10.1093/bioinformatics/btp324

1034 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009.
1035 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.
1036 doi:10.1093/bioinformatics/btp352

1037 Lindsay SJ, Rahbari R, Kaplanis J, Keane T, Hurles ME. 2019. Similarities and differences in patterns
1038 of germline mutation between mice and humans. *Nat Commun* **10**:1–12. doi:10.1038/s41467-
1039 019-12023-w

1040 Mak SST, Gopalakrishnan S, Carøe C, Geng C, Liu S, Sinding MHS, Kuderna LFK, Zhang W, Fu S,
1041 Vieira FG, Germonpré M, Bocherens H, Fedorov S, Petersen B, Sicheritz-Pontén T, Marques-
1042 Bonet T, Zhang G, Jiang H, Gilbert MTP. 2017. Comparative performance of the BGISEQ-500
1043 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience* **6**:1–

13. doi:10.1093/gigascience/gix049

Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol* **2**:1940–1955. doi:10.1038/s41559-018-0717-x

Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, Skov L, Belling K, Theil Have C, Izarzugaza JMG, Grosjean M, Bork-Jensen J, Grove J, Als TD, Huang S, Chang Y, Xu R, Ye W, Rao J, Guo X, Sun J, Cao H, Ye C, van Beusekom J, Espeseth T, Flindt E, Friborg RM, Halager AE, Le Hellard S, Hultman CM, Lescai F, Li S, Lund O, Løngren P, Mailund T, Matey-Hernandez ML, Mors O, Pedersen CNS, Sicheritz-Pontén T, Sullivan P, Syed A, Westergaard D, Yadav R, Li N, Xu X, Hansen T, Krogh A, Bolund L, Sørensen TIA, Pedersen O, Gupta R, Rasmussen S, Besenbacher S, Børglum AD, Wang J, Eiberg H, Kristiansen K, Brunak S, Schierup MH. 2017. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**:87–91. doi:10.1038/nature23264

Martin HC, Batty EM, Hussin J, Westall P, Daish T, Kolomyjec S, Piazza P, Bowden R, Hawkins M, Grant T, Moritz C, Grutzner F, Gongora J, Donnelly P. 2018. Insights into Platypus Population Structure and History from Whole-Genome Sequencing. *Mol Biol Evol* **35**:1238–1252. doi:10.1093/molbev/msy041

Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. 2017. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun* **8**:15183. doi:10.1038/ncomms15183

Patch AM, Nones K, Kazakoff SH, Newell F, Wood S, Leonard C, Holmes O, Xu Q, Addala V, Creaney J, Robinson BW, Fu S, Geng C, Li T, Zhang W, Liang X, Rao J, Wang J, Tian M, Zhao Y, Teng F, Gou H, Yang B, Jiang H, Mu F, Pearson J V., Waddell N. 2018. Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. *PLoS One* **13**:e0190264. doi:10.1371/journal.pone.0190264

Pfeifer SP. 2017. Direct estimate of the spontaneous germ line mutation rate in African green monkeys. *Evolution (N Y)* **71**:2858–2870. doi:10.1111/evo.13383

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA Van der, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178. doi:10.1101/201178

Prasad A, Lorenzen ED, Westbury M V. 2021. Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *bioRxiv* 2021.03.03.433733. doi:10.1101/2021.03.03.433733

1076 Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki S Al, Dominiczak A, Morris
1077 A, Porteous D, Smith B, Stratton MR, Consortium U, Hurles ME. 2016. Timing, rates and
1078 spectra of human germline mutation. *Nat Genet* **48**:126–133. doi:10.1038/ng.3469

1079 Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, Chen BJ, Kher M,
1080 Banks E, Ames DC, English AC, Li H, Xing J, Zhang Y, Matise T, Abecasis GR, Salerno W,
1081 Zody MC, Neale BM, Hall IM. 2018. Functional equivalence of genome sequencing analysis
1082 pipelines enables harmonized variant calling across human genetics projects. *Nat Commun*
1083 **9**:4038. doi:10.1038/s41467-018-06159-4

1084 Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, Consortium W, Wilkie AO, McVean G, Lunter
1085 G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in
1086 clinical sequencing applications. *Nat Genet* **46**:912–918. doi:10.1038/NG.3036

1087 Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman
1088 N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. 2010. Analysis of genetic
1089 inheritance in a family quartet by whole-genome sequencing. *Science* **328**:636–639.
1090 doi:10.1126/science.1186802

1091 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011.
1092 Integrative genomics viewer. *Nat Biotechnol*. doi:10.1038/nbt.1754

1093 Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013.
1094 Characterizing and measuring bias in sequence data. *Genome Biol 2013 145* **14**:1–20.
1095 doi:10.1186/GB-2013-14-5-R51

1096 Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, Quinlan AR. 2019. Large, three-
1097 generation human families reveal post-zygotic mosaicism and variability in germline mutation
1098 accumulation. *Elife* **8**:e46922. doi:10.7554/eLife.46922

1099 Ségurel L, Wyman MJ, Przeworski M. 2014. Determinants of mutation rate variation in the human
1100 germline. *Annu Rev Genomics Hum Genet* **15**:47–70. doi:10.1146/annurev-genom-031714-
1101 125740

1102 Smeds L, Qvarnström A, Ellegren H. 2016. Direct estimate of the rate of germline mutation in a bird.
1103 *Genome Res* **26**:1211–1218. doi:10.1101/gr.204669.116

1104 Straube D, Juen A. 2013. Storage and shipping of tissue samples for DNA analyses: A case study on
1105 earthworms. *Eur J Soil Biol* **57**:13–18. doi:10.1016/J.EJSOBI.2013.04.001

1106 Tatsumoto S, Go Y, Fukuta K, Noguchi H, Hayakawa T, Tomonaga M, Hirai H, Matsuzawa T, Agata
1107 K, Fujiyama A. 2017. Direct estimation of de novo mutation rates in a chimpanzee parent-
1108 offspring trio by ultra-deep whole genome sequencing. *Sci Rep* **7**:1–12. doi:10.1038/s41598-
1109 017-13919-7

1110 Thomas GWC, Wang RJ, Nguyen J, Alan Harris R, Raveendran M, Rogers J, Hahn MW. 2021.
1111 Origins and Long-Term Patterns of Copy-Number Variation in Rhesus Macaques. *Mol Biol Evol*
1112 **38**:1460–1471. doi:10.1093/MOLBEV/MSAA303

1113 Thomas GWC, Wang RJ, Puri A, Rogers J, Radivojac P, Hahn MW, Thomas GWC, Wang RJ, Puri
1114 A, Harris RA, Raveendran M. 2018. Reproductive longevity predicts mutation rates in primates.
1115 *Curr Biol* **28**:1–5. doi:10.1016/j.cub.2018.08.050

1116 Tian S, Yan H, Kalmbach M, Slager SL. 2016. Impact of post-alignment processing in variant
1117 discovery from whole exome data. *BMC Bioinformatics* **17**:403. doi:10.1186/s12859-016-1279-z

1118 Tiley GP, Poelstra JW, Reis M Dos, Yang Z, Yoder AD. 2020. Molecular Clocks without Rocks:
1119 New Solutions for Old Problems. *Trends Genet* **2020**. doi:10.1016/j.tig.2020.06.002

1120 Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, Kronenberg ZN, Hormozdiari F,
1121 Raja A, Pennacchio LA, Darnell RB, Eichler EE. 2017. Genomic Patterns of De Novo Mutation
1122 in Simplex Autism. *Cell* **171**:710-722.e12. doi:10.1016/J.CELL.2017.08.047

1123 Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, Raja A, Baker C,
1124 Hoekzema K, Stessman HA, Zody MC, Nelson BJ, Huddleston J, Sandstrom R, Smith JD,
1125 Hanna D, Swanson JM, Faustman EM, Bamshad MJ, Stamatoyannopoulos J, Nickerson DA,
1126 McCallion AS, Darnell R, Eichler EE. 2016. Genome Sequencing of Autism-Affected Families
1127 Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* **98**:58–74.
1128 doi:10.1016/J.AJHG.2015.11.023

1129 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T,
1130 Shakir K, Roazen D, Thibault J, Banks E, Garimella K V, Altshuler D, Gabriel S, DePristo
1131 Geraldine A Van der Auwera MA. 2013. From FastQ data to high confidence variant calls: the
1132 Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma* **11**:10.
1133 doi:10.1002/0471250953.bi1110s43

1134 Venn O, Turner I, Mathieson I, De Groot N, Bontrop R, McVean G. 2014. Strong male bias drives
1135 germline mutation in chimpanzees. *Science* **344**:1272–1275. doi:10.1126/science.344.6189.1272

1136 Wang RJ, Radivojac P, Hahn MW. 2021a. Distinct error rates for reference and nonreference

1137 genotypes estimated by pedigree analysis. *Genetics* **217**:1–10.
1138 doi:10.1093/GENETICS/IYAA014

1139 Wang RJ, Raveendran M, Harris RA, Murphy WJ, Lyons LA, Rogers J, Hahn MW. 2021b. De novo
1140 mutations in domestic cat are consistent with an effect of reproductive longevity on both the rate
1141 and spectrum of mutations. *bioRxiv* 2021.04.06.438608. doi:10.1101/2021.04.06.438608

1142 Wang RJ, Thomas GWC, Raveendran M, Harris RA, Doddapaneni H, Muzny DM, Capitanio JP,
1143 Radivojac P, Rogers J, Hahn MW. 2020. Paternal age in rhesus macaques is positively
1144 associated with germline mutation accumulation but not with measures of offspring sociability.
1145 *Genome Res* **30**:826–834. doi:10.1101/gr.255174.119

1146 Wilson Sayres MA, Makova KD. 2011. Genome analyses substantiate male mutation bias in many
1147 species. *BioEssays* **33**:938–945. doi:10.1002/bies.201100091

1148 Wingett S. 2017. Illumina Patterned Flow Cells Generate Duplicated Sequences. *QC Fail Artic about*
1149 *common next-generation Seq Probl*.

1150 Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, Baker R, Thach DC,
1151 Iyer RK, Vockley JG, Niederhuber JE. 2016. New observations on maternal age effect on
1152 germline de novo mutations. *Nat Commun* **7**:1–10. doi:10.1038/ncomms10486

1153 Wu FL, Strand AI, Cox LA, Ober C, Wall JD, Moorjani P, Przeworski M. 2020. A comparison of
1154 humans and baboons suggests germline mutation rates do not track cell divisions. *PLOS Biol*
1155 **18**:e3000838. doi:10.1371/journal.pbio.3000838

1156 Yang C, Zhou Y, Marcus S, Formenti G, Bergeron LA, Song Z, Bi X, Bergman J, Rousselle MMC,
1157 Zhou C, Zhou L, Deng Y, Fang M, Xie D, Zhu Y, Tan S, Mountcastle J, Haase B, Balacco J,
1158 Wood J, Chow W, Rhie A, Pippel M, Fabiszak MM, Koren S, Fedrigo O, Freiwald WA, Howe
1159 K, Yang H, Phillippy AM, Schierup MH, Jarvis ED, Zhang G. 2021. Evolutionary and
1160 biomedical insights from a marmoset diploid genome assembly. *Nat 2021 5947862* **594**:227–
1161 233. doi:10.1038/s41586-021-03535-x

1162 Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-BLAST: a tool to
1163 design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**:134.
1164 doi:10.1186/1471-2105-13-134

1165