File S1

Supplementary Text

2nd Generation Sequencing

Blood and tissues were obtained in agreement with protocols reviewed and approved by the Gibbon Conservation Center. DNA was extracted from blood or cell lines, and paired-end libraries were prepared with the Illumina TruSeq chemistry. Libraries were sequenced on the HiSeq 2000 platform, generating 2x100 bp reads. Four different sequencing centers contributed sequence data (**Table S8**). Multiple runs were performed to generate a minimum of 10X mean coverage on each sample after all processing. Mean coverage ranged from 11.5X to 19.5X (**Table S9)**. Exome capture using the TruSeq Exome Enrichment Kit (Illumina) was performed on one NLE sample (Vok, 116x coverage) and one SSY sample (Monty, 64x coverage).

Read Mapping and Variant Calling

Sequences in FASTQ format were trimmed with cutadapt (MARTIN 2011) to remove Illumina TruSeq adapter sequences. Reads with less than 25 nucleotides left after trimming were dropped, along with their mates. The remaining reads were aligned to nomLeu1 with Stampy (v. 1.0.17) (LUNTER and GOODSON 2011). For the two *N. leucogenys* (NLE) samples, Stampy was used in its "hybrid mode" where alignment with BWA (v. 0.5.9) (LI and DURBIN 2009) is attempted first. A substitution rate of 0.001 was specified, along with BWA minimum seed length of 2, fraction of missing alignments 0.0001, and quality threshold 10. For the non-NLE samples, stampy was used with a substitution rate of 0.015 (KIM *et al.* 2011). Local realignment at indel sites was performed with the Genome Analysis Toolkit (GATK, v. 1.4-37) (MCKENNA *et al.* 2010; DEPRISTO *et al.* 2011). PCR duplicates were then removed with samtools. Picard (v. 1.70) (http://sourceforge.net/projects/picard/) CleanSam was run on the output. The two samples from each genus were then merged using Picard MergeSamFiles, and the resulting files were split using samtools (LI *et al.* 2009) into 100 files containing ~180 contigs each to facilitate further parallel processing. The GATK UnifiedGenotyper was run and Single Nucleotide Variants (SNVs) and indels with a quality score of at least 50 were retained to create a mask of variant sites to be excluded from base quality score recalibration (BQSR). The BQSR steps were run with the standard set of covariates, and the resulting files were merged across all samples. The GATK indel realignment tools were then run again to standardize alignment of indels across the samples. Default settings were used except that "BadCigar" reads were excluded and BAQ calculation was added. The UnifiedGenotyper from GATK version 2.1-11 was then used to call SNVs and indels in each genomic part using the "EMIT_ALL_SITES" mode (with the BAQ calculation included) to produce VCF files with data for all genomic positions. (Version 2.1 was used for this step to allow multiallelic calling). VCFs for all genomic parts were then merged

using a custom perl script. Annotations were added to specify the consensus quality score of the nomLeu1 reference sequence at each position.

Exome sequencing data was processed separately from the shotgun data but using the same bioinformatic pipeline. The exome targeted regions were lifted over to the nomLeu1 genome using the UCSC liftOver utility with the default parameters, and the emit-all VCF of the exome capture data were restricted to these target loci. Sites within these loci with less than 30x coverage or over 200x coverage for any sample were also removed, while corresponding sites in the whole genome data with a variant quality less than 20 were called as homozygous reference in all samples.

Finding Accurately Called Segregating Sites

Machine learning classification techniques, such as variant quality score recalibration, have been successfully used to find a subset of sites that are predicted to be truly segregating in a sample. However, the authors know of no technique that has been used to predict whether or not individual genotypes have been correctly called, and as such downstream methods that presume that the genotypes are correct when they are in fact incorrect may suffer accordingly. To this end we developed a ML classification protocol to find a set of segregating sites where every genotype within is predicted to be correct for use in our Principal Components Analysis. Broadly, this protocol uses the comparison of the whole genome sequencing (WGS) and whole exome sequencing (WES) truth set to train several largely disparate classifiers. The classifiers are then used to predict the accuracy of individual genotypes across the genome. We note that this protocol may introduce some level of bias with respect to the agglomerative properties of sites (owing to the increased difficulty in calling heterozygous vs. homozygous genotypes) as opposed to individual genotypes, and as such this approach would be undesirable for evaluating, say, the site frequency spectrum.

More specifically, the machine learning (ML) suite Weka version 3.6.8 (HALL *et al.* 2009) was used to classify the whole genome genotype data at all called segregating sites, with the aim of finding a subset of very high quality sites. Using the definition of "correct" from our profiling of errors, we collected the set of all genotypes that were incorrectly called in the genome, and a random and equally sized sampling of genotypes that were called correctly for both our NLE and our non-NLE (SSY) sample. The following features were used in the machine learning analysis: approximate read depth, the next-best genotype likelihood, the haplotype score, the read-position bias score, the base quality rank score, the total mapping-quality 0 reads, the root-mean square mapping quality, the fraction of reads spanning deletions, the probability of strand bias, mapping quality rank sum test, quality by depth, the maximum likelihood expectation of the allele counts and allele frequency, the quality of the reference base, whether the call is from the NLE or the non-NLE sample, and the combined p-value of the

distribution of read depths observed at the site. All but the last three features were taken directly from the GATK output. For the last feature, 2-tailed p-values of the read depth observed at a site for an individual were taken, based on that individual's empirical distribution of read depths, and these p-values were combined across individuals using the Method of Fisher to give a single description of read-depth for the site.

Using the features described above we generated a training set and evaluated the performance of a variety of classifiers using 10-fold cross validation. Four techniques – multilayer perceptron, ridor, rotation forest and classification by regression– showed reasonable performance (75%-85% accuracy). However, as our goal is to find genotypes that have been correctly called, we used cost-sensitive classification to minimize our false discovery rate. Using a simple grid search, we found a cost-matrix that maximized each classifier's positive-predictive value by down-weighting the relative cost of false negatives versus false positives. 10-fold cross validation gave the estimates of accuracy and positive predictive value shown in **Table S10**.

To reduce overfitting, each iteration from the 10-fold cross validation for each learner type (e.g. rotation forest) was kept, with each of the folds being a function of the least-significant digit in the SNP position (e.g., Rotation Forest$_2$ would be trained on all SNP positions where 2 is not its least-significant digit, and tested on all SNP positions where 2 is its least-significant digit). These 4x10 learners were then used in the assessment of genotype accuracy in our exome data. We then classified a genotype call as correct if all four classifiers predicted that the genotype was correct, and we classified a site as correct if all genotypes at a site were classified as correct. To increase our sample size of genotypes, the ML included sites that would have been masked out in the CNV calls (which represent less of a problem given that our measure of correctness is really a metric of consistency). Our final assessment of accuracy, however, included the CNV masks. Over a total of 54,528 sites that are segregating in the WES (after applying our filters) and marked as being genotyped correctly according to the 4 learners above, there was a total of 1 genotyping error and this error occurred in our non-NLE sample.

Approximate Bayesian Computation analysis

Sequence divergence essentially reflects an upper bound for when populations split and can give a false signal of the phylogeny if the time of coalescence for sequences can fall within the ancestral population of the extant populations of interest (DEGNAN and ROSENBERG 2006). Therefore in order to investigate the gibbon phylogeny at the population divergence level we applied a Bayesian coalescent-based method that explicitly take into account sequence and population divergence simultaneously. Most methods that currently perform this task such as BEAST (DRUMMOND and RAMBAUT 2007) are not suited to large datasets that result from 2nd generation sequencing. Therefore we have developed an Approximate Bayesian Computation (ABC) (BEAUMONT *et al.* 2002) method that can cope with large amounts of sequence data, is not dependent on haplotype phase and can

incorporate information derived from our modeling of errors from comparing WGS with high coverage WES data. We aimed to use the ABC framework to a) identify the most likely species topology for the four gibbon genera that underwent WGS and b) estimate key parameters of the gibbon speciation process (specifically effective population sizes and divergence times).

Methods

Data: ABC analysis was performed on two data sets. For the first, to approximate independence among regions we identified loci consisting of 1kb of total callable sequence separated by at least 50kb. In addition to the masks and coverage filters described in the main manuscript we also masked CpG consistent sites as well as conserved phastCons elements inferred from primate genomes with a further 100bp padding either side of the element. Loci were then identified that were 50kb from the nearest exon and where the 1,000 callable bases fall within a maximum of 3kb of contiguous nomLeu1 reference sequence (i.e. callable bases are not necessarily contiguous) (see **Fig S1** for a cartoon of the distribution of these loci). This resulted in 12,413 1kb loci (total of ~12Mb). Because these loci are relatively distant from each other (>50kb apart) inter-locus linkage can be ignored and as they are relatively short (max 3kb) intra-locus recombination should be negligible. Therefore we do not incorporate recombination parameters into our simulations, only mutation plus the demographic parameters of interest. However because of the large number of loci analyzed, our data will approach the analytical expectations of the coalescent and thus should allow accurate and precise estimates of the correct model and associated parameters.

In addition we generated a set of 11,323 200bp loci under the same criteria except the loci were orientated to lie on a known exon (i.e. genic versus non-genic loci) with the allowance of a maximum of 100bp either side of the known exon boundary, spanning a total of 4kb with a minimum of 1kb separating any two loci. This relatively small latter distance will likely violate the assumption of independent genealogies between loci somewhat but increasing this distance in to 5kb severely reduced the number of loci, which will decrease accuracy and precision more readily. The choice of 200bp per loci for genic regions was motivated by the average length of exons in the gibbon genome of 213bp. Variant sites were polarized against the aligned human reference genome, hg19, using the multiz 11-way alignments from UCSC.

Phylogeny Models and Parameter Priors: We treat all possible phylogenetic relationships amongst the four gibbon genera as distinct models (we also treat the two species within *Hylobates* as one population to reduce the model space). Therefore we need to consider a total of 15 models describing the population divergence relationship among the 4 genera, 12 asymmetric (**Fig S15**) and 3 symmetric (**Fig S16**). We also considered an instantaneous 4-way hard polytomy in a second ABC model testing analysis. As is standard for coalescent-based phylogenetic approaches the models are described by two classes of parameters,

mean nucleotide diversity, θ, and branch lengths in units of expected number of substitutions, τ. Given an estimate of the mean mutation rate, μ, the former can be transformed into an estimate of N_e using $\theta = 4N_e\mu$ and the latter can be transformed into a divergence time in generations, *t*, using $\tau = t\mu$. Priors ranged between 0.0001-0.03 for all θ and τ parameters. Unless otherwise stated all prior distributions for all demographic parameters are all uniformly distributed on a $\log_{10}$ (x) scale. The justification for this prior range is that, assuming a mutation rate of 1×10^{-8} per site per year and a 10 year generation time for gibbons, our individual priors are equivalent to a time of divergence of 100kya-30mya and an N_e of 2,500-750,000. These ranges take into account the uncertainty we have with regard ape speciation times and ancestral diversity. Thirty million years sits at the upper end of the range when apes are thought to have diverged from other old world monkeys (ZALMOUT *et al.* 2010). The earliest known "sub-species" split times observed in great apes is ~80kya (western and cross river gorilla), while the earliest known "species" split time (which is what our gibbon data essentially is) is 175kya (western and eastern gorillas), with most being much older (on the order of millions of years) (PRADO-MARTINEZ *et al.* 2013). Similarly, other estimates of great ape heterozygosity range from ~0.0005-0.0025, with ancestral θ estimates based on pairwise sequentially markovian coalescent (PSMC) analysis not exceeding 0.005, while we observed in a PSMC analysis of the same Gibbon samples used here (CARBONE *et al.* 2014) that the ancestral N_e is unlikely to have risen to values greater than 50,000.

When estimating parameters from the best model we included separate HPI and HMO populations with their own θ values and a new τ parameter for their divergence time. In addition we included a version of this model where the four genera split simultaneously, and thus only incorporate two ancestral θ parameters (one for the HPI and HMO ancestral population and one for the ancestral population of all four genera) and two τ parameters. Finally the analysis with this latter model was repeated using true uniform priors (rather than $\log_{10}$ transformed priors) for the two τ parameters (see **Results** in main manuscript for more details on this analysis)

Simulations: Coalescent simulations of the 8 individuals (16 chromosomes) were performed using a version of ms (Hudson) modified for Python that allowed fast parallel processing. In total we performed 10^6 random draws of the parameter space and simulated a θ-scaled genealogy for each locus. In order to account for mutation rate heterogeneity across loci we estimated relative sequence divergence for all loci, taking the average sequence divergence for each of the eight gibbon individuals from hg19. These individual locus estimates were then normalized around a mean of 1, allowing us to follow the approach of Rannala and Yang (RANNALA and YANG 2003) and scale θ for each individual locus in our demographic simulations.

Stochastic Error Modeling: We used the error profiles for the singleton and non-singleton categories described above in Vok and Monty to construct an error model $E = \langle S, M \rangle$ for a particular sample that could transform perfectly correct data generated by coalescent simulations into data reflective of the error processes that are likely to have occurred during whole genome sequencing and post processing. We found that with our bioinformatic pipeline the total number of observed singletons was always less than or equal to the true number. Therefore S was calculated as the proportion of missing singletons, or the probability of not calling a true singleton in the WGS data. During a coalescent simulation of genetic data S reflects the rate at which true singletons will be hidden or dropped and the genotype called as homozygous reference. To construct M we took the 3x3 confusion matrix generated for non-singletons and divided the number in each element of the matrix by the sum of all elements within their respective columns. During a simulation of genetic data, for any site not classed as a true singleton but still segregating, the values within a particular column of M reflect the probabilities of a multinomial distribution that determines the rate that a true genotype of a particular type will be transformed to one of the two other genotypes or stay the same.

To apply our error correction to a) non-exome regions in the two target samples and b) non-exome regions in the other six samples for which there was no WES we constructed separate E models for each read depth ≥ 7X (i.e, we constructed $E_i = \langle S_i, M_i \rangle$, where E_i is the estimated error at read-depth i). Singleton calling was markedly better in the reference taxon (S~99%), then in the non-reference taxon (S~96%). For S_i, error rates initially decreased up to ~20x but past 30 showed substantial increases in errors, presumably from uncalled CNVs (**Fig S17**). Similar to our findings with singletons, WGS/WES discordance rates initially decreased with increasing read depth, but from read depths of ~20x onwards discordance rates again began to increase (**Fig S18**). Given the error profiles observed with respect to coverage, we chose to break our error rate estimations into three read-depth classes; 7-20x, 21-29x and ≥30x. For the first class, we assumed that our per-read-depth estimates were correct, and for the last class, consistent with our assumption from the WES data, that the WGS calling was perfect. For the middle class, however, we conservatively assumed a constant error rate taken from the average error rate from read-depths 18-20.

This information allowed us to construct an overall E model for a particular sample, regardless of whether it was one of the two target samples or not, by taking a weighted average of E_i, with weights determined by the empirical distribution of read depths at the specific regions of interest for sites between 7x and that individual's 95th percentile of read depth. The E_i models for Vok and Monty were used for NLE and non-NLE samples respectively to take into account any mapping biases. We assumed errors were uncorrelated between individuals. As the error models were generated with respect to the nomLeu1 reference (rather than some ancestral reference) we simulated an additional haploid NLE sample to orient the error correction

appropriately. Summary statistics were generated from the simulations for both with and without stochastically modeling error processes in order to examine the affect of the former.

Ancestral state misidentification adjustment: 2% ancestral state misidentification was incorporated into simulations by calculating the expected number of sites to experience a mutation along the hg19 lineage for each locus (1000bp x 2% = 20 sites). The number of sites to actually "flip" (i.e. assign the wrong ancestral state) for each locus during a simulation is drawn from a Poisson distribution with this mean. These sites are then randomly assigned to positions along the locus, though only positions that are found to segregate amongst the gibbon chromosomes need to be flipped.

Summary Statistics: We computed the following summary statistics to describe the data for every pair of populations: number of shared derived polymorphisms across loci, number of private derived polymorphisms in each population and the number of private fixed sites in each population **(Table S11)**. These statistics are known to contain substantial information about population demography (WAKELEY 1996) and are utilized in the program MIMAR (BECQUET and PRZEWORSKI 2007). These statistics are particularly useful in the case of short read sequence data as they do not require haplotype inference. We use the mean of these summary statistics across all loci to describe the data (unlike MIMAR where these summaries are used to calculate a likelihood of the data for each locus individually, which is computationally intensive for the amount of data considered here). We also explored the use of the variance of these same summary statistics across loci but found they added little to our ability to infer parameters in the model while contributing more noise to the partial least squares (PLS) transformation and reducing the proportion of correctly inferred simulated topologies using simulated data (see below). Other summary statistics that might traditionally be considered useful for demographic inference such as Tajima's D were not utilized due to the small sample size for each species. Therefore our method is unable to infer parameters such as population growth rates.

Inference: ABC analysis was performed using two different regression adjustments depending on their application. When estimating model parameters we utilized ABCtoolbox (WEGMANN *et al.* 2010), which implements a general linear model (GLM) adjustment (LEUENBERGER and WEGMANN 2010) on retained simulations. The GLM adjustment, by modeling the parameters as the predictor rather than the response variable, avoids one particular limitation of the standard linear regression adjustment of Beaumont et al. (BEAUMONT *et al.* 2002) where the posterior distribution can end up being non-zero in parameter space that actually lies outside the prior bounds. To maximize sufficiency but limit dimensionality, the full set of summary statistics was transformed into PLS components (WEGMANN *et al.* 2009) and we used the change in Root Mean Square Error (RMSE) to guide

the choice of number of components. These PLS components were then used to estimate parameters. In our analysis to assess the ability of our ABC framework to determine the correct species topology/model (see below) we compared the marginal distributions across models using both the GLM adjustment above and the multinomial logistic regression (LR) method previously described by Fagundes et al. (FAGUNDES *et al.* 2007). The performance of the two methods was generally very similar though the LR method demonstrated a slight increase (~3%) in the proportion of correctly recovered models. Because of this slightly better performance, added to the fact the LR method is by far the most popular regression adjustment method used for ABC model choice (CSILLÉRY *et al.* 2010) and we are less concerned in this case with extrapolating the posterior distributions outside of the prior range (as we are using categorical classes and all classes have equal prior probability), we chose to use this method for all subsequent analysis using an adapted version of the R function calmod.r as. However the use of either method is likely to give very similar results in our particular framework. 1% of simulations were retained for the GLM (parameter estimation) and LR (model choice) adjustments. PCA was used for comparing the multidimensional distribution of summary statistics using the ‘‘prcomp’’ function in R.

Using simulated data to assess the ability to determine the correct species topology

In order to assess the reliability of our method to infer the correct species tree from a set of alternatives we simulated 10,000 random pseudo-observed datasets from our model and demographic parameter priors and attempted to recover the true topology using the ABC machinery. We explored which combination of summary statistics most often inferred the correct topology and found that the six summary statistics describing the mean number of shared sites for a pair of populations was most effective. Adding more summary statistics (such as mean pairwise fixed or private differences or the variance of the number of shared sites across loci) reduced the proportion of correctly inferred simulated topologies and thus were discarded for this analysis.

Using the LR method for the error-corrected non-genic data we recovered the correct model 88.4% (7,989/9,042) of the time (for 958 topologies the LR method failed to converge), the correct model was one of the top 3 models 99.1% (8,959/9,042) of the time and had a posterior probability greater than 5% 98.3% (8,894/9,042) of the time. Using a more naïve method (the direct method, DR) of the proportion of retained simulations from each model (PRITCHARD *et al.* 1999) we recovered the correct model 77.6% (7,757/1,000) of the time, the correct model was one of the top 3 models 96.7% (9,673/10,000) of the time and had a posterior probability greater than 5% 99.5% (9,950/10,000) of the time. For the 958 occasions when the LR method failed, the DR method inferred the correct model 792 (83.0%) times and was within the top 3 models on every occasion bar one, with a minimum posterior probability of 0.07. This suggests the failure of the LR method to

converge results from either complete separation or because all the retained simulations are only from one model, rather than an inability to detect the correct model. The posterior probabilities using both the LR and DR methods were highly informative with regard to the correct model. However the LR method generally demonstrated a better level of discrimination between the true model and all other models (**Fig S4**). Therefore we decided to use this method for all subsequent analysis. For the uncorrected data, there was a slight increase in the ability to infer the correct topology (which is unsurprising given the error model essentially adds noise) where, for example, using the LR method we recovered the correct model 8,298/9,048 (91.7%) of the time (the 952 topologies for the LR method failed). The proportion of occasions where we inferred the correct topology for corrected genic data was similar (86.9%).

In order to obtain a better idea of where our method failed (and where it performed well) we performed a more targeted set of simulations and again attempted to infer the correct model using our ABC framework. We first chose the total height of the four taxa species tree (*Tanc*) to be one of three values (in units of mutations per site): 0.01, 0.005 and 0.001. Assuming a mutation rate of $1x10^{-9}$ per site per year this is equivalent to 10, 5 and 1 million years. We then chose the ϑ values across the tree to be either fixed at 0.001 in all present and past populations, or for the present values to be 0.0012, 0.0004, 0.002, 0.0008 (thus roughly reflecting present day estimated ϑ for gibbons in this study) and for the ancestral populations to reflect the combinations of these ϑ values (i.e. 0.0012 + 0.0004 = 0.0016, 0.0016 + 0.002 = 0.0036, 0.0036 + 0.0008 = 0.0044, thus the N_e gets increasingly bigger going back in time to increase the probability of incomplete lineage sorting in the ancestral populations, i.e. we make the problem "harder"). The purpose of the latter set of ϑ values is not to choose values that necessarily reflect reality (though we attempt to pick sensible choices that will prove intuitively useful), but to examine how the method tolerates changes in θ compared to utilizing fixed values (which should be an "easier" problem). Finally we simulated either an asymmetrical tree or symmetrical tree. Thus there are 12 parameter combinations representing 12 scenarios that define our simulations. For each of these 12 scenarios we choose the two most recent divergence times (*Tanc* is the third and last event and is already set) over a range as follows.

1) The most recent divergence event is chosen to be equal to *Tanc*/α, where α varies from 1.1-5, with steps of 0.1. The smaller α, the closer the most recent divergence event will be to the final divergence event. For example for α of 1.1, when *Tanc* = 0.001 this means that the most recent divergent event occurs at 0.0009, which results in a separation time of only ~100,000 in years.
2) The second divergence event is chosen based on β, which ranges between 0.01-0.99 in steps of 0.01, with the value of β reflecting the distance from the final divergence event as a percentage of the time between the first and last

divergence event. Again, a smaller β reflects a second divergence event that is very close to the last divergence event (conversely a high β reflects a divergence event that is very close to the first divergence event).

We used our ABC machinery to determine the posterior probabilities for the true model under these 12 scenarios while varying α and β along a two dimensional grid. Posterior probabilities were determined using either the DR or LR method. **Fig S5** shows these results with the grid of α and β being the X and Y axis and the surface of the posterior probabilities across this grid shown in the Z-axis. It is immediately apparent that the LR produces much higher posterior probabilities compared to the DR method, with much of the surface of the former being at or close to 1.0. In both cases reducing *Tanc* to 0.001 reduces the posterior probability, but the affect is markedly worse for the DR method. Unsurprisingly, for the asymmetric model our ability to infer the true model is best when α is highest and β is 0.5 and for the symmetric model when α is highest and β is 1. This essentially reflects situations where the divergence events are maximally separated in terms of branch lengths. The LR method performs particularly well in most cases, with the posterior probability only decreasing markedly at the edges of the grid (and for *Tanc* 0.001 and 0.005 the value of α has almost no effect) suggesting that for a *Tanc* realistic for gibbons (>4Mya), the method will only perform sub optimally if the second divergence event is very close to the first or last divergence event, and even then the posterior probability will still likely be one of the higher values across all 15 possible topologies. Varying θ across populations (annotated as fixed in **Fig S5**) does not appear to have a large effect for the DR or LR methods but does appear to further exacerbate the poorer performance of the method when *Tanc* = 0.001, where we presume incomplete lineage sorting becomes particularly prevalent such that it obscures true tree topology.

Using simulated data to assess the effect of stochastically modeling error processes on parameter inference

To assess how stochastically modeling errors in our simulations for ABC analysis are likely to affect inference of parameters, we applied our entire analysis framework to simple demographic scenarios where the data was simulated to mimic errors in next generation sequencing that occur as a result of variable per site coverage and read-specific base miscalling. Specifically we assessed the affect of our analysis strategy for estimating parameters under two demographic scenarios (**see Fig S19**):

Scenario A. We estimate θ in a one-population of constant size model. We sample two individuals (four chromosomes) from the population, one of which is used as a target sample for which we know the true genotypes to generate an *E* model, as well as a reference chromosome.

Scenario B. We estimate τ in a two-populations of constant size with divergence model, where the θ values in the two present day populations and the ancestral population are fixed at 0.001, 0.002 and 0.001 respectively. We sample two individuals from

each population (eight chromosomes), with one from each used as a target sample for which we know the true genotypes to generate an *E* model. The reference genome is drawn from the first population.

Our framework for these analyses for a particular demographic model (with a specifc parameter value of interest) involves four primary stages (further details are given below):

1. Simulating medium coverage next generation sequence data in "exome regions" under the demographic model of interest and calling genotypes from this data, which along with the known true genotypes allows the construction of E_i models for target samples at individual read depths ≥ 7X.
2. Simulating medium coverage next generation sequence data in "neutral regions" under the demographic model of interest and calling genotypes from this data. This is essentially our observed data for the ABC analysis.
3. Constructing overall *E* models for all individuals based on their simulated empirical distribution of read depths at the neutral regions and the E_i from the target samples in stage 1.
4. Performing an ABC analysis to infer parameters where we stochastically introduce errors into the ABC Monte Carlo simulations based on the *E* model constructed for each sample.

In theory we should apply all steps for every parameter value we would like to explore under a given demographic scenario. We aimed to examine scenarios A and B for θ and τ values that ranged from 0.0001 to 0.01 in steps of 0.1 $\log_{10}$ units. This would involve applying our framework 21 times for each scenario. Even for this relatively modest exploration of the parameter space, this would be particularly time consuming for stage 4 where we must generate hundreds of thousands of Monte Carlo simulated datasets each time. However we have found that error profiles to generate *E* models are only minimally affected by the specific θ or τ used in our two particular demographic scenarios (they become much more variable in more complex demographic scenarios, data not shown) (**Fig S20**). Therefore, in practice we construct our read depth specific E_i and overall *E* models (steps 1 and 3) and generate Monte Carlo simulations (step 4) only for θ or τ = 0.001 (i.e. we utilize the midpoint parameter value) and always apply the same empirical distribution of read depths for each sample for neutral regions regardless of the value of θ or τ (step 2). Therefore, we only need to perform steps 1, 3 and 4 once across all θ or τ values for a particular demographic scenario, though it is still necessary to perform step 2 to generate the observed data for each of the individual 21 parameter values. This will mean our ABC analysis is only optimally applied to the parameter value of 0.001 and all other inference will be slightly sub-optimal compared to how the method could be applied in practice.

Simulating the truth set: For a given demographic model and parameter value of interest we simulated 20,000 x 8 exons each of length 150bp for the appropriate number of chromosomes given the model (including a reference chromosome) using ms. This gives a total diploid sequence length of 24MB per individual, approximating the amount of data generated by most WES capture kits. This simulated sequence data reflects the true genotypes and essentially represent the high coverage WES data from Vok and Monty in our analysis of real data.

Simulating next generation sequence data in exomic regions: To simulate medium coverage next generation sequencing data for these same true underlying genotypes, we assume each individual was sequenced to a mean coverage of 10X and each site for each individual is assigned some number of reads (i.e. coverage) by randomly drawing from a poisson distribution with λ=10. For each true heterozygous site the number of reads "sequenced" for each allele is drawn from a binomial distribution with p=0.5. Each read at a site is assigned a Phred-scaled quality score, Q, from a truncated poisson distribution with λ=30 such that Qs are repeatedly drawn until a value ≤40 is obtained (i.e. Q is limited to 40) and an error is introduced at a rate proportional to this Q value (for example for reads with Q=30 there is a 1 in 100 probability that it will be assigned the wrong base call). In keeping with our use of the infinite sites model in *ms* and to simply our downstream analysis we limit bases to only two types, reference and non-reference, rather than he four bases A, C, T and G.

In reality there is considerably more complexity with regard to how coverage and base calling error is distributed across the genome than is considered in our framework. Coverage is necessarily correlated at sites because reads span at least 100bp of sequence and paired end reads are used in most circumstances, while sequencing errors tend to be more frequent at the end of reads. In addition Q values assigned by Illumina sequencing software are frequently not truly representative of the true error rate and are base pair and context dependent. In addition there may also be differences in the proportion of reads that correctly map to the reference genome when utilizing a mixture of reference on non-reference species individuals. We also do not consider the effect of Indels or repeats and CNVs, which can introduce additional error from misalignment.

Calling genotypes from the simulated next generation sequence data: We recoded the maximum likelihood genotype and Bayesian variant calling algorithms described in Depristo et al.(DEPRISTO *et al.* 2011) to only consider two alleles and called genotypes (assuming a heterozygosity parameter of 0.001) from our simulated next generation sequence data for any sites with coverage ≥7x. Sites with a variant quality value <40 were assigned as homozygous reference in all samples. These called genotypes essentially represent the medium coverage WGS data in our error modeling.

Construction of E_i for target samples: Given these simulated WGS and truth data sets we can construct E_i for target samples as described for the real data. When examining error profiles for both scenarios we see a consistent pattern of a decreased proportion of missing singletons with increased coverage, with values of ~15% missing singletons at 7X but rising to no errors by 20X (**Fig S20**). This trend is largely in line with our real data (**Fig S17**), suggesting that our framework is capturing the most important error processes, though the real data is much noisier which is likely due to many of the additional factors described above.

Simulating next generation sequence data in neutral regions: To simulate "neutral regions" (the observed data) we use the same simulation framework described above to introduce errors and generate called genotypes for the exomic regions except that we simulate 12,000 1kb regions (to mimic our real data of 1kb regions). *E* models are then constructed for each sample given a simulated distribution of coverage in the neutral regions.

ABC inference: We then perform two separate ABC analyses, one with the introduction of stochastic errors via the *E* model and one without. For Scenario A we use the mean number of segregating sites per locus as the summary statistic and for Scenario B we use the mean number shared, private and fixed sites per locus between the two populations.

Results: The framework described above (subjective to some simplifications to aid tractability) was applied for a range of θ and τ values under scenarios A and B respectively. We then compared our estimated parameter values (using the mode and median of the posterior distributions) to the true simulated value. The use of the *E* model consistently improves the estimate of parameters, with the effect being particularly noticeable for larger value of τ (**Fig S6**). For example when the true τ is 0.010, modeling errors results in almost no difference with the estimate τ (τ = 0.0098) compared to when errors are ignored (τ = 0.0079). In units of time in years this is a difference of ~2my. The RMSE for θ and τ when using the *E* model is 36% and 8% of that respectively when not using the *E* model. Thus our simulations suggest that stochastic modeling of error processes in ABC simulations can improve the inference of parameters for 2nd generation sequencing data.

G-PhoCS analysis

The Markov Chain Monte Carlo (MCMC) Bayesian coalescent-based method described by Gronau et al. (GRONAU *et al.* 2011) was performed using the software G-PhoCS to estimate θ and τ values for a bifurcating tree (we ignored the effect of migration). On this occasion we included a human haploid sequence (hg19) as an outgroup for the overall gibbon phylogeny. The same 12,431

1kb loci and assumed bifurcating tree from the ABC analysis described above were utilized and the mutation rate was fixed individually for each locus as above using the normalized divergence values. The gamma prior for θ was set to be relatively broad and the same for all present and ancestral populations with shape, α, = 2 and rate, β, =1,000. Gamma priors for τ were also set to be relatively broad, with the α value always 2. However, either *a)* β was set as 200 for all τ values such that the mean (α/β) = 0.01, which, when assuming μ = 1x10-9 /site/year equates to 10My or *b)* individual β values were set for each τ such that the mean value reflected rough estimates from the ABC analysis or for the human/gibbon split time from Carbone et al. (CARBONE *et al.* 2014) (**Table S1**). Starting values for the MCMC chain for each parameter were chosen randomly from the interval of 0.8-1.2 * these mean values. Preliminary runs under *b)* were used to tune the MCMC mixing (as this is the multidimensional parameter space that is likely to be most important for estimating parameters in this case and mixing properties can change in different parts of the space), such that the rate of acceptance was between 20%-70% for all parameters of the model.

Once we obtained good mixing properties we ran three independent MCMC chains for both prior settings *a)* and *b)* for a total of six chains. We allowed 10,000 samples as burn-in followed by 100,000 samples for estimating parameters (this sample size should be large enough that we do not require independent samples to get unbiased estimates due to correlation among consecutive samples). The Markov chain converged to stationarity much quicker than the utilized burn-in period, and all six runs converged to the same tationary distribution, though prior setting *a)* required slightly more samples as the starting positions were further away from the converged τ values. Results were processed using the software Tracer (http://tree.bio.ed.ac.uk/software/tracer/).

References

BEAUMONT M. A., ZHANG W., BALDING D. J., 2002 Approximate Bayesian computation in population genetics. Genetics **162**: 2025–2035.

BECQUET C., PRZEWORSKI M., 2007 A new approach to estimate parameters of speciation models with application to apes. Genome Res **17**: 1505–1519.

CARBONE L., HARRIS R. A., GNERRE S., VEERAMAH K. R., 2014 Gibbon genome and the fast karyotype evolution of small apes. Nature.

CSILLÉRY K., BLUM M., GAGGIOTTI O. E., AL E., 2010 Approximate Bayesian computation (ABC) in practice. Trends in ecology & ….

DEGNAN J. H., ROSENBERG N. A., 2006 Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genet **2**: e68.

DEPRISTO M. A., BANKS E., POPLIN R., GARIMELLA K. V., MAGUIRE J. R., HARTL C., PHILIPPAKIS A. A., DEL ANGEL G., RIVAS M. A., HANNA M., MCKENNA A., FENNELL T. J., KERNYTSKY A. M., SIVACHENKO A. Y., CIBULSKIS K., GABRIEL S. B., ALTSHULER D., DALY M. J., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43**: 491–498.

DRUMMOND A. J., RAMBAUT A., 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **7**: 214.

FAGUNDES N. J. R., RAY N., BEAUMONT M., NEUENSCHWANDER S., SALZANO F. M., BONATTO S. L., EXCOFFIER L., 2007 Statistical evaluation of alternative models of human evolution. Proceedings of the National Academy of Sciences **104**: 17614–17619.

GRONAU I., HUBISZ M. J., GULKO B., DANKO C. G., SIEPEL A., 2011 Bayesian inference of ancient human demography from individual genome sequences. Nat Genet **43**: 1031–1034.

HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P., WITTEN I. H., 2009 The WEKA data mining software. SIGKDD Explor. Newsl. **11**: 10.

KIM S. K., CARBONE L., BECQUET C., MOOTNICK A. R., LI D. J., DE JONG P. J., WALL J. D., 2011 Patterns of Genetic Variation Within and Between Gibbon Species. Molecular Biology and Evolution **28**: 2211–2218.

LEUENBERGER C., WEGMANN D., 2010 Bayesian Computation and Model Selection Without Likelihoods. Genetics **184**: 243–252.

LI H., DURBIN R., 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754–1760.

LI H., HANDSAKER B., WYSOKER A., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G., DURBIN R., 1000 GENOME PROJECT DATA PROCESSING SUBGROUP, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

LUNTER G., GOODSON M., 2011 Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res **21**: 936–939.

MARTIN M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: pp. 10–12.

MCKENNA A., HANNA M., BANKS E., SIVACHENKO A., CIBULSKIS K., KERNYTSKY A., GARIMELLA K., ALTSHULER D., GABRIEL S., DALY M., DEPRISTO M. A., 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20**: 1297–1303.

PRADO-MARTINEZ J., SUDMANT P. H., KIDD J. M., LI H., KELLEY J. L., LORENTE-GALDOS B., VEERAMAH K. R., WOERNER A. E., O'CONNOR T. D., SANTPERE G., CAGAN A., THEUNERT C., CASALS F., LAAYOUNI H., MUNCH K., HOBOLTH A., HALAGER A. E., MALIG M., HERNANDEZ-RODRIGUEZ J., HERNANDO-HERRAEZ I., PRÜFER K., PYBUS M., JOHNSTONE L., LACHMANN M., ALKAN C., TWIGG D., PETIT N., BAKER C., HORMOZDIARI F., FERNANDEZ-CALLEJO M., DABAD M., WILSON M. L., STEVISON L., CAMPRUBÍ C., CARVALHO T., RUIZ-HERRERA A., VIVES L., MELÉ M., ABELLO T., KONDOVA I., BONTROP R. E., PUSEY A., LANKESTER F., KIYANG J. A., BERGL R. A., LONSDORF E., MYERS S., VENTURA M., GAGNEUX P., COMAS D., SIEGISMUND H., BLANC J., AGUEDA-CALPENA L., GUT M., FULTON L., TISHKOFF S. A., MULLIKIN J. C., WILSON R. K., GUT I. G., GONDER M. K., RYDER O. A., HAHN B. H., NAVARRO A., AKEY J. M., BERTRANPETIT J., REICH D., MAILUND T., SCHIERUP M. H., HVILSOM C., ANDRÉS A. M., WALL J. D., BUSTAMANTE C. D., HAMMER M. F., EICHLER E. E., MARQUES-BONET T., 2013 Great ape genetic diversity and population history. Nature **499**: 471–475.

PRITCHARD J. K., SEIELSTAD M. T., PEREZ-LEZAUN A., FELDMAN M. W., 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution **16**: 1791–1798.

RANNALA B., YANG Z., 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics **164**: 1645–1656.

WAKELEY J., 1996 Distinguishing Migration from Isolation Using the Variance of Pairwise Differences. Theoretical Population Biology **49**: 369–386.

WEGMANN D., LEUENBERGER C., EXCOFFIER L., 2009 Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood. Genetics **182**: 1207–1218.

WEGMANN D., LEUENBERGER C., NEUENSCHWANDER S., EXCOFFIER L., 2010 ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics **11**: 116.

ZALMOUT I. S., SANDERS W. J., MACLATCHY L. M., GUNNELL G. F., AL-MUFARREH Y. A., ALI M. A., NASSER A.-A. H., AL-MASARI A. M., AL-SOBHI S. A., NADHRA A. O., MATARI A. H., WILSON J. A., GINGERICH P. D., 2010 New Oligocene primate from Saudi Arabia and the divergence of apes and Old World monkeys. Nature **466**: 360–364.