Examining Phylogenetic Relationships among Gibbon Genera using Whole Genome Sequence Data Using An Approximate Bayesian Computation Approach

Krishna R. Veeramah*, August E. Woerner*, Laurel Johnstone*, Ivo Gut†, Marta Gut†, Tomas Marques-Bonet‡, Lucia Carbone**, Jeff D. Wall§§, Michael F. Hammer*

*Arizona Research Laboratories Division of Biotechnology, University of Arizona, Tucson, AZ, USA
§Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, USA
†CNAG (Centro Nacional de Analisis Genomico), Baldiri Reixac 4, 08028 Barcelona, Spain
‡ICREA at Insitit de Biologia Evolutiva (CSIC/UPF), Dr. Aiguader 88, 08003 Barcelona, Spain
**Department of Behavioral Neuroscience, Oregon Health and Science University, Portland, OR, USA
§§Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA
Running Title:

ABC analysis of Gibbon Genomes

Corresponding author name and address:

Michael F Hammer
ARL Division of Biotechnology, Keating Building 111K, University of Arizona, Tucson, AZ 85721, USA

E-mail: mfh@email.arizona.edu

Phone: (520)-621-9228

Fax: (520)-621-9247
Abstract

Gibbons are believed to have diverged from the larger great apes ~16.8 Mya and today reside in the rainforests of Southeast Asia. Based on their diploid chromosome number, the family Hylobatidae is divided into four genera, Nomascus, Symphalangus, Hoolock and Hylobates. Genetic studies attempting to elucidate the phylogenetic relationships among gibbons using karyotypes, mtDNA, the Y chromosome, and short autosomal sequences have been inconclusive. To examine the relationships among gibbon genera in more depth, we performed 2nd generation whole genome sequencing to a mean of ~15X coverage in two individuals from each genus. We developed a coalescent-based Approximate Bayesian Computation method incorporating a model of sequencing error generated by high coverage exome validation to infer the branching order, divergence times, and effective population sizes of gibbon taxa. Although Hoolock and Symphalangus are likely sister taxa, we could not confidently resolve a single bifurcating tree despite the large amount of data analyzed. Instead our results support the hypothesis that all four gibbon genera diverged at approximately the same time. Assuming an autosomal mutation rate of 1x10^{-9}/site/year this speciation process occurred ~5 Mya during a period in the Early Pliocene characterized by climatic shifts and fragmentation of the Sunda shelf forests. Whole genome sequencing of additional individuals will be vital for inferring the extent of gene flow among species after the separation of the gibbon genera.
Introduction

The family *Hylobatidae*, commonly known as gibbons, are believed to have diverged from the larger great apes ~16.8 My (CARBONE et al. 2014). Sometimes known as small apes, gibbons demonstrate substantial morphological differentiation from the great apes; their much smaller bodies are highly adapted to an arboreal mode of locomotion in the rainforests of Southeast Asia. They also demonstrate very little sexual dimorphism that may, in part, be related to their generally monogamous mating patterns (FUENTES 2000) (although some gibbon species develop differences in coat color at sexual maturity).

Each species demonstrates distinct ‘call’ and ‘song’ types (GEISSMANN 2002); however, attempts to classify gibbon species and genera based solely on morphological features have been problematic (MOOTNICK 2006). Primarily on the basis of their karyotypes, gibbons are now divided into four major genera, with *Nomascus*, *Symphalangus*, *Hylobates* and *Hoolock* each possessing 52, 50, 44, and 38 diploid chromosomes, respectively. While many genetic studies have been performed, including a number based on karyotypes (MÜLLER et al. 2003), mitochondrial DNA (HAYASHI et al. 1995; TAKACS et al. 2005; MONDA et al. 2007; WHITTAKER et al. 2007; VAN NGOC et al. 2010; MATSUDAIRA and ISHIDA 2010), Y chromosomes (CHAN et al. 2012), ALU repeats (MEYER et al. 2012), and short stretches of autosomal sequence (KIM et al. 2011; WALL et al. 2013), the phylogenetic relationships among the four gibbon genera remain unresolved, with at least seven different topologies being supported by different data.

A recent study examined ~1.5 Mb of orthologous autosomal sequence generated by 2nd generation sequencing from one individual representing each of the four genera (WALL et al. 2013). This study, too, was inconclusive and suggested that the gibbon genealogy demonstrates substantial incomplete lineage sorting (ILS). However, the experimental design was limited by
the lack of a suitable reference genome (short reads were aligned to highly divergent human hg19 assembly). To examine the species tree relationships among gibbons, as well as estimate key demographic parameters such as the time when the various gibbon genera diverged, we generate whole genome sequence data from eight individuals representing all four gibbon genera and utilize the newly released gibbon (nomLeu1) reference genome (CARBONE et al. 2014) for mapping and variant calling. Then we apply a coalescent-based Approximate Bayesian Computation (ABC) approach that can handle large amounts of sequence data and that corrects for potential sequencing error and reference genome mapping bias.

Materials and Methods

2nd Generation Sequencing

Blood and tissues were obtained in agreement with protocols reviewed and approved by the Gibbon Conservation Center. More details on all aspects of the methods are provided in the Supplementary Information. DNA was extracted from blood or cell lines, and paired-end libraries were prepared with the Illumina TruSeq chemistry. Libraries were shotgun sequenced on the HiSeq 2000 platform, generating 2x100 bp reads. Multiple runs were performed to generate a minimum of 10X mean coverage on each sample after all post-processing. Mean coverage ranged from 11.5X to 19.5X. Exome capture using the TruSeq Exome Enrichment Kit (Illumina) was also performed on one NLE sample (Vok, 116x coverage) and one SSY sample (Monty, 64x coverage)

Read Mapping and Variant Calling
Trimmed reads from the shotgun sequencing 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 removed with samtools. GATK UnifiedGenotyper was run separately on the two samples from each genus 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. The GATK indel realignment tool was run again to standardize alignment of indels across all samples. UnifiedGenotyper from GATK version 2.1-11 (to allow multiallelic calling) was used to produce a final set of SNVs and indels. Each site was annotated with the consensus quality score of the nomLeu1 reference sequence. Exome sequencing data was processed separately from the shotgun data but using the same bioinformatic pipeline (more details can be found in the Supplementary Text).

**Masking of the Gibbon Reference Genome for Downstream Analysis**

The nomLeu1 genome is composed of 17,968 contigs, ranging in size from 2,496 bases to ~74 MB. As small loci may be compressed and represent duplications in the gibbon genome that have not been properly separated during the assembly process, we masked out all scaffolds less than 1MB in length, yielding 273 scaffolds that span ~2.73 GB. UCSC’s gibbon-human pairwise alignments were used to identify non-autosomal sequence. Specifically, gibbon loci that aligned
to human X, Y or M in UCSC’s “net” alignments (KENT et al. 2003) were masked, along with locations in the gibbon genome that were not primary alignments to locations in the human genome. Further, locations where the gibbon reference quality was below a phred-quality of 50, repeats (identified by Tandem Repeat Finder (BENSON 1999) or by RepeatMasker (SMIT et al. 1996)), LAVA elements identified in Carbone et al. (CARBONE et al. 2014), CNVs with an estimated ploidy >2.5 in any sample (also identified in Carbone et al. (CARBONE et al. 2014), infinite sites violations, positions where any sample has less than 7x coverage, or more than their 95th percentile read depth, and bases within 3bp of any indel called were excluded, unless otherwise specified, from downstream analysis.

**Profiling of Sequencing Errors**

As our whole genome sequencing (WGS) coverage is ~15X per sample, it is likely that some observed genotype calls will not reflect the true underlying genotypes, either because of a combination of low coverage and errors in the sequencing reads or as a result of choices in the bioinformatic processing (e.g., variant quality score thresholds). Therefore, we compared genotype calls between the WGS and whole exome sequencing (WES) data at the same genomic positions to profile errors present in our medium coverage WGS, assuming high coverage WES reflects a “truth set”. Separate profiles were constructed for Vok and Monty. These profiles were then used a) as training data to find a set of high confidence variable SNPs in all eight gibbon samples using machine learning (ML) methods, and b) to stochastically model error processes in our subsequent ABC analysis.

Our profiling and modeling of errors made the following simplifying assumptions: (1) after masking, any WES site with 30x <= coverage <= 200x is called without error and reflects
the truth dataset and all other bases are ignored, (2) per site read-depth and mapping bias (which we naively model by noting whether the sample belongs to the same taxon as the reference or not) account for all genotyping errors observed, and (3) all false negatives (i.e., SNPs present in the WES, but not present in the whole genome data) are singletons. We profiled errors between the WES truth set and WGS for a given target sample (Vok or Monty) via two categories: errors involving singleton polymorphisms (defined with respect to the nonLeu1 reference), and genotyping errors when the polymorphism is segregating with a non-reference allele present in two or more chromosomes.

For the former category, we recorded the number of singleton calls in our WGS data versus the number in the WES truth set. For a site to be considered a singleton, a single non-reference allele must be present in the either the WGS, the WES truth set or both for the target sample, and it must not have been observed in any other sample. Singleton sites that agree and disagree between the WGS and WES truth sets are considered as ‘correct’ and ‘incorrect’, respectively.

For the latter category (i.e., any site that is segregating in either the WGS, the WES truth set or both but is not defined as a singleton as described above) we created 3x3 confusion matrices over the set of genotype calls {Reference homozygous, Heterozygous, Alternative homozygous} to describe all 9 possible WGS vs WES truth set genotype calls for the target sample. The diagonals of these confusion matrices reflect sites with concordant calls between the WGS vs WES truth set (i.e., they are ‘correct’) and off-diagonals represent discordance, and thus potential errors (i.e., ‘incorrect’). For example, the sum of the middle column will represent all heterozygous sites in the WES truth set. The middle element of this column represents sites that were also called as heterozygous in the WGS data (i.e., correctly called sites). The top (bottom)
element represents a genotyping error in the WGS data where the site is truly heterozygous but was called as homozygous reference (alternate).

**Finding Accurately Called Segregating Sites**

ML 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 an 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 WGS and 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 ML suite Weka version 3.6.8 (Hall et al. 2009) was used to classify the WGS 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. A variety of features from the GATK output (see **Supplementary Text** for the entire
list) as well as 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 were used in the ML analysis. Using the various features 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). After various optimization procedures, we 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. PCA was performed using smartpca (PATTERSON et al. 2006) and visualized using R.

**ABC analysis**

Our ABC framework was designed 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) (more detail can be found in the Supplementary Text).

Data: ABC analysis was performed on two data sets containing independent loci of small enough length such that intra and interlocus recombination could reasonably be ignored. Set 1 included 12,413 non-genic loci consisting of 1 kb of total callable sequence across a contiguous stretch of no more than 3 kb separated by at least 50 kb and at least 50 kb from the nearest exon. Set 2 included 11,323 genic loci consisting of 200 bp of total callable sequence across a contiguous stretch of no more than 4 kb separated by at least 1 kb (this distance will likely violate our assumption of independence but increasing this distance substantially decreased the number of
usable loci and thus reduced the accuracy and precision of our inference to a greater extent, with an allowance of a maximum of 100 bp of the locus lying adjacent to an exon and the rest lying in the exon (Fig S1). In addition to the masks and coverage filters described above, we also masked CpG consistent sites as well as conserved phastCons (SIEPEL et al. 2005) elements inferred from primate genomes with a further 100 bp padding either side of the element. Variant sites were polarized against the aligned human reference genome, hg19.

Phylogeny Models and Parameter Priors: We treated all possible phylogenetic relationships among the four gibbon genera as distinct models (including, where applicable, the true polytomy model). The models are described by two classes of parameters, mean population nucleotide diversity, θ, and branch lengths, τ, in units of expected number of substitutions (thus mutation rates per site per generation do not need to be explicitly stated during the analysis). Priors ranged between 0.0001-0.03 for all θ and τ parameters (a justification for these priors ranges is given in the supplementary material). Unless otherwise stated all prior distributions for all demographic parameters (θ and τ) are all uniformly distributed on a log_{10} (x) scale.

Simulations: Coalescent simulations of demographic models and parameters were performed using a version of ms (HUDSON 2002) modified for Python that allowed fast parallel processing to allow us to efficiently simulate the thousands of loci seen in our observed data. To account for mutation rate heterogeneity among 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 $E = <S, M>$ 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 $\geq 7X$ (i.e, we constructed $E_i$, the estimated error rate at a particular read-depth $i$). This 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. The $E_i$ models for Vok and Monty were used for NLE and non-NLE samples respectively to take into account any potential mapping biases.

**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 (1000 bp 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 a position along the locus with equiprobability, though only positions that are found to segregate amongst the gibbon chromosomes need to be flipped computationally.

**Summary Statistics:** We computed the following summary statistics to describe the observed and simulated data for every pair of populations across all loci: mean number of shared derived polymorphisms, mean number of private derived polymorphisms in each population and the mean number of private fixed sites in each population. We also explored including the variance of these summary statistics across all 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 reduced the proportion of correctly inferred simulated topologies using simulated pseudo-observed data.

**Inference:** We used the logistic regression (LR) method previously described (FAGUNDES et al. 2007) to perform model choice. 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. Before ABC analysis for parameter inference 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. 1% of simulations were retained for the GLM (parameter estimation) and LR (model choice) adjustments.

**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 $\theta$ and $\tau$ 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 (rather than just to infer the ancestral state as in the ABC analysis). The same 12,431 1 kb loci and the bifurcating species tree with the highest posterior probability 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 $\theta$ was set to be relatively broad and the same for all present and ancestral populations with shape, $\alpha$, = 2 and rate, $\beta$, =1,000. Gamma priors for $\tau$ were also set to be relatively broad, with the $\alpha$ value always 2. However, either a) $\beta$ was set as 200 for all $\tau$ values or b) individual $\beta$ values were set for each $\tau$ such that the mean value reflected rough estimates from the ABC analysis or the human/gibbon split time from Carbone et al. (Carbone et al. 2014) (Table S1). We ran three independent MCMC chains for both prior settings a) and b). We allowed 10,000 samples as burn-in followed by 100,000 samples for estimating parameters. The Markov chain converged to stationarity much
quicker than the utilized burn-in period, and all six runs converged to the same stationary
distribution. Results were processed using the software Tracer
(http://tree.bio.ed.ac.uk/software/tracer/).

Results

*2nd generation sequencing and validation*

We performed 2nd generation WGS on two individuals (one male and one female) from each of
the four gibbon genera (*Table 1*). For our *Nomascus* samples, represented by the species
*leucogenys* (NLE, the northern white-cheeked gibbon), the two individuals examined differed
from the (NCBI Project 13975 GCA_000146795.1) nomLeu1 reference genome. For our
*Hylobates* samples (the most diverse genus with ~13 species), we examined one individual each
from the *H. moloch* (HMO, Javan gibbon) and *H. pileatus* (HPI, Pileated gibbon). Our
*Symphalangus* sample is represented by two individuals from the species *syndactylus* (SSY,
Siamang gibbon). It is important to point out that the two *Hoolock* samples from the *leuconedys*
species (HLE, Eastern hoolock gibbon) represent the only wild born individuals present in the
study, whereas all other individuals were captive-born (i.e., offspring of individuals living in
zoos). We also mention that matings between different gibbon species (and even different
genera) are known to result in viable offspring in captivity (*MYERS* and *SHAFER* 1979;
*MOOTNICK* 2006; *HIRAI* *et al.* 2007). If any of the individuals in our sample are indeed hybrids
between different species, our analysis may be affected in unexpected ways.

After post-processing the sequence data we obtained a mean coverage of 15X (min =
11.5X, max =19.5X) (*Fig S2*). As previous work has indicated a relatively high divergence
between gibbon genera, we attempted to incorporate potential reference bias into our post-
processing by utilizing a higher substitution rate (1.5%) when mapping sequence reads for non-NLE samples, and by using a hybrid mapper, Stampy (Lunter and Goodson 2011) to increase sensitivity. To validate our variant calling we performed high coverage WES on one NLE individual and one non-NLE sample (the male SSY sample). Mean coverage for WES data was 116X (compared with 14X for WGS data) and 64X (compared with 13X for WGS data), respectively. Human-based exome capture has been shown to be effective in primates as diverged from humans as macaques (Jin et al. 2012). Utilizing only exome calls with coverage between 30X and 200X we found slightly greater concordance between the WGS and WES data for the NLE (99.6%) versus non-NLE samples (99.4%) (Table S2). Noticeably when only examining singleton variants, calling was markedly better in the reference taxa (~99% of exome-called sites identified in the WGS data) than in the non-reference taxa (~96%), suggesting reference biases may still exist in our data for rare variants in non-reference taxa.

**Genetic diversity among gibbon genera**

Within genera diversity, assessed for this dataset by Carbone et al. (Carbone et al. 2014), demonstrated that NLE samples had the highest level of nucleotide diversity (π~2.2x10⁻³), while values as low as ~7.3x10⁻⁴ were observed in the HPI sample. Nucleotide diversity for the HMO sample was also relatively high at ~1.7x10⁻³, followed by SSY (~1.4x10⁻³), and then the two wild born HLE (~8x10⁻³). By way of comparison, π ranges from approximately 0.5-1.0x10⁻³ in humans, 1.8x10⁻³ in western lowland gorillas, and 2.3x10⁻³ in Sumatran orangutans (Prado-Martinez et al. 2013). To examine the relative levels of genetic differentiation among the gibbon genera we performed PCA on the individual samples. For this analysis we examined di-allelic SNPs called in all individuals. High-quality SNPs were identified by using concordance
with the WES data to train a ML algorithm to predict highly confident genotype calls across the whole genome and in samples that did not undergo WES. In addition, to ensure independence of SNPs, we randomly selected sites that were separated by at least 100 kb when on the same scaffold. This resulted in a dataset of 25,531 high quality genome-wide independent SNPs. The first four principal components accounted for 40.2%, 31.2%, 24.6% and 3.5% of the variation, respectively (Fig S3A and S3B). The four genera showed substantial genetic differentiation and were clearly separated in the PCA plot in the first two components, though no clear inter-genera phylogenetic relationship emerged. Individuals from the same species showed high similarity suggesting limited inter-genera hybridization or contamination. The two Hylobates species could be clearly distinguished in PC4. We were also able to reproduce the same patterns when only using a random subset of ~200 SNPs (Fig S3C and D), suggesting it may be possible to perform relatively low coverage shotgun sequencing from a number of different gibbon species and use a similar approach to this in order to identify a small yet powerful set of species specific SNPs. This could be particularly important for management of gibbons in zoos when it can often be difficult to distinguish different species or even genera based on fur alone, sometimes leading to accidental hybrids.

**A Coalescent-based ABC Analysis of the Gibbon Phylogeny**

Unless species branch lengths are several orders of magnitude larger than the expected time to the most recent common ancestor of sequences within a species, it is important to model stochasticity in the distribution of gene trees across loci when inferring an underlying species tree (ROSENBERG and NORDBORG 2002). Current Bayesian coalescent-based methods such as BEAST (DRUMMOND and RAMBAUT 2007) that explicitly take into account sequence and
population divergence simultaneously to infer species trees are generally computationally intractable for large datasets (Bryant et al. 2012). Therefore, in order to infer the species topology for gibbon genera we developed an Approximate Bayesian Computation (ABC) (Beaumont et al. 2002) method for inference of a species tree with four taxa. The method can also infer species divergence times and effective population sizes for a given topology, can handle large amounts of sequence data, is not dependent on haplotype phase, and incorporates information derived from our modeling of errors from comparing WGS with high coverage WES data.

Analogous to the Bayesian approach of Gronau et al. (Gronau et al. 2011), which uses an analytical derivation to determine the likelihood of the full data given typical population genetic parameters, the data required for this ABC method are short, independent loci as we assume no intra-locus recombination and free recombination between loci. The latter is a necessary convenience given that no recombination map is currently available for gibbons. Thus, we assembled a set of independent ‘non-genic’ sequences that mapped at least 50 kb away from genes (~12,000 1 kb loci) and that excluded CpG consistent sites as well as evolutionarily conserved elements (Siepel et al. 2005) (Fig S1). Mutations detected in these loci are expected to represent neutral variation and to evolve at a relatively constant rate. To reduce reference-mapping bias, we also assembled an analogous set of independent ‘genic’ loci that span exons (~11,000 200 bp loci) and that should have lower diversity, recognizing that these loci may have been subjected to natural selection, which may bias any parameter estimates.

Analysis of pseudo observed data generated by simulations demonstrated that we were able to detect the correct topology from randomly drawn datasets using our method 88.4% of the time, with the correct model among the three highest posterior probabilities 99% of the time.
Analysis of a more targeted set of pseudo observed data demonstrated that the method is only likely to fail when an internal branch is extremely small (almost instantaneous in evolutionary terms) or when the total height of the tree in units of expected number of substitutions is on the order of 0.001 (equivalent to about 1 million years for apes) \textbf{(Fig S5)}, which is unrealistic for gibbons.

As most ABC analyses are based on performing simulations to approximate an otherwise analytically intractable likelihood function, we also attempted to stochastically model sequence errors (missing singletons and incorrect genotype calls at other segregating sites) that are likely to have occurred in the real 2\textsuperscript{nd} generation sequencing data. Errors were introduced into coalescent simulations by an $E$ model constructed by comparing the WGS with the high coverage WES data. By incorporating this $E$ model we found through simulated pseudo observed data that we could infer more accurate estimates of $\theta$ and $\tau$ under very simple demographic scenarios (one population with a constant size $\theta$ through time and two populations of constant size that diverged at some time, $\tau$, in the past) \textbf{(Fig S6)}. A full description of the above validation of our ABC framework using pseudo observed data is given in the \textbf{Supplementary Information}.

Prior to the ABC analysis of the real data we examined the one-dimensional distribution for each individual summary statistic from 10,000 random simulations from the $\theta$ and $\tau$ parameter space and found a good fit to our non-genic and genic observed data, while a PCA also demonstrated a good multidimensional fit \textbf{(Fig S7)}.

\textbf{Table 2} shows the posterior probabilities from the ABC analysis for all phylogenetic models for the observed data for both the non-genic and genic loci using the corrected (with stochastic errors introduced via the $E$ model) and uncorrected coalescent simulations (a total of
four analyses). No topology dominates the analysis, with three to four topologies having posterior probabilities >10% in the corrected simulations. The best topology using non-genic and genic loci for the corrected simulations differ, and both still maintain relatively low posterior probabilities of 19%. Two topologies appear most prominent with posterior probabilities >10% in all four analyses and the highest means across all four analyses and both (genic and non genic) corrected analyses. One is the most frequently observed topology in the sequence divergence analysis (((SSY, HLE), NLE),(HPI,HMO)) of Carbone et al. (CARBONE et al. 2014) and the other is a related topology where (HPI, HMO) and NLE are swapped as the most external groups with HLE and SSY remaining as sister taxa. Together the posterior probability for both these related topologies sum to 30-32%. However, in general the posterior probabilities are lower than typically observed in our pseudo observed datasets suggesting that we have little confidence in the true topology. This is consistent with the hypothesis of a rapid radiation of gibbon species from a large ancestral population.

The simplest phylogenetic description of this process would be a four-way hard polytomy. Therefore we constructed an additional model with all four genera diverging at the same time and analyzed this scenario within the same ABC framework as the previously examined 15 bifurcating topologies (i.e., we examined 16 different models in total). This did not affect our ability to infer the correct model using pseudo-observed data sets. As with considering only strictly bifurcating topologies we were able to detect the correct topology from randomly drawn datasets from all 16 models 87.3% of the time. Of the 16 individual models the instantaneous model was the one with the lowest proportion of correctly predicted pseudo-observed data sets but was still high at 82.4%.
Despite having the lowest predictive value, when we examined the real data the posterior probability for instantaneous model ranged from 87-90% for both the non-genic and genic loci and for the corrected and uncorrected simulations (Table 3). The posterior probabilities of the 15 bifurcating topologies after the addition of the instantaneous model were necessarily much lower but still highly correlated with the previous values with $r^2$ ranging between 0.91-0.98. Thus, our ABC analysis strongly supports a relatively instantaneous hard polytomy for the divergence of the four gibbon genera over that of a particular bifurcating topology. Evidence for this can also be seen visually by examining a PCA of the summary statistics for 1,000 random datasets from each of the 16 models, with the instantaneous model lying within the center of the cloud of all models and the observed data found firmly within this part of the cloud (Supp Fig 8). Other polytomy combinations may also fit the data (for example a model with the initial divergence of three lineages, followed by a later Hoolock and Symphalangus divergence) but our ability to reliably discriminate such additional intermediate models is likely to further worsen given our instantaneous model already shows reduced predictive ability compared to the other fully bifurcating models.

Estimation of Parameters describing Gibbon Demography

To estimate when this rapid radiation may have taken place we constructed a model where all four genera diverge simultaneously with the addition of a subsequent divergence of the two Hylobates species. This resulted in a model with seven $\theta$ and two $\tau$ parameters. The summary statistics from the non-genic loci were transformed into PLS components to infer these parameters. Posterior distributions and parameter estimates are shown in Table S3 and Fig. S9. These results are based on 15 PLS components, the value at which the largest reduction in the
RMSE was observed across all parameters, **Fig. S10**, and for which the CIs for $\tau$ were considered relatively reliable based on how often the true value fell within the estimated 95% CIs using 1,000 pseudo observed datasets (Veeramah et al. 2012).

Observed values of $\pi$ described above were within the 95% CI for the $\theta$ values estimated by the ABC analysis for present-day species and showed the same relative pattern with the highest value in the NLE and lowest value in the HPI sample. The divergence time, $\tau_1$, for the two *Hylobates* samples was ~50% less than that for the divergence time of the four gibbon genera, $\tau_2$, which is consistent with the relative difference in sequence divergence of ~50% seen in Carbone et al. (Carbone et al. 2014). Because the priors were log$_{10}$ scaled, the associated 95% CI values potentially could be larger in absolute values (i.e. $10^{\text{val}}$) than if the observed posterior distribution had been shifted towards a smaller branch length. Therefore, we re-ran the ABC analysis using un-scaled flat priors for the two $\tau$ values, which resulted in highly similar median values but much narrower 95% CIs (**Table 4**, **Table S4**, **Fig. S11**). We note that these CIs were somewhat anticonservative as assessed by pseudo observed datasets (see column “HDPI 95% fit” of **Table 4** and **Table S4**). When we assume a $\mu$ of $1\times10^{-9}$ per site per year * 3/4 (to take into account that we excluded CpG sites (Hodgkinson and Eyre-Walker 2011)) this results in an estimate for the time of the gibbon radiation of $1.6+3.5 = 5.1$ Mya ($\tau_1$,$\tau_2$ combined limits of 95% CI 2.5-7.7 Mya) and a split time of 1.6 Mya (95% CI 0.6-2.9 Mya) for the two *Hylobates* samples. In addition, assuming 10 years per generation for gibbons (Harvey et al. 1987) and thus a $\mu$ of $7.5\times10^{-9}$ per generation, $N_e$ for extant species varies from 57,000 (NLE) to 7,500 (HPI). Interestingly, the ancestral gibbon $N_e$ is estimated to be much larger at 132,000 (107,000-162,000) (**Fig. 1a**) as would be expected under a model of substantial ILS. It should be noted that the estimate of the ancestral *Hylobates* population size (based on $\theta_{T1}$) may be
somewhat unreliable as the regressed posterior distribution shows a major shift from the raw retained posterior distribution (Fig. S11) while this was also the θ value for which the largest number of PLS components were needed to obtain a reasonable reduction in the RMSE (Fig. S12).

One potential source of error in estimating parameters is ancestral state misidentification due to back mutations along the human lineage, which was used as an outgroup (HERNANDEZ et al. 2007). Our simulated data assumed an infinites sites model. Assuming a human-gibbon split time of 16.8 Mya and μ of 1x10⁻⁹ per site per year, each site has ~98% chance ((1-1x10⁻⁹)^(16,800,000)) of not experiencing a substitution along the human branch. Therefore, we conducted the ABC parameter estimation on a set of 10⁵ simulations where we incorporated a 2% rate of random ancestral allele misidentification. Though this binary model of back mutation is highly simplistic (e.g., it does not take into account mutations to another base pair type or trinucleotide context), we found it had only minimal impact on our 95% CIs compared with the same number of simulations that did not incorporate some ancestral state misidentification error (Table S5). This suggests that our divergence time estimates may be only slightly underestimated by not accounting for this error.

To investigate the effect of imposing a model of instantaneous speciation rather than bifurcating species divergence on our parameter inference, we also modeled the five gibbon species assuming the best sequence phylogeny from Carbone et al. (CARBONE et al. 2014) and that was also suggested by our ABC model choice analysis, (((SSY,HLE)NLE)(HPI,HMO)) (Table S6, Fig. 1b, Fig. S13, Fig. S14). The median estimates of the posterior distributions for the seven θ parameters common to both the bifurcating and instantaneous models (five extant population values as well θ_T1 and θ_anc) were similar, while the 95% CI for θ_T1 and θ_T2 were broad
and uninformative. Consistent with the rapid speciation hypothesis (even when allowing bifurcating speciation), $\tau_2$, $\tau_3$, $\tau_4$ was roughly equivalent to $\tau_2$ for the instantaneous speciation model, with $\tau_3$ and $\tau_4$ being an order of magnitude smaller (i.e., very short internal branch lengths).

We also applied the 1 kb data to the method of Gronau et al. (Gronau et al. 2011) as this approach is based on a similar model (i.e., the coalescent with population divergence) as our bifurcating ABC analysis; however, it should be more accurate for parameter estimation as it is based on an exact model likelihood rather than an approximation (although it does not currently incorporate the possibility of sequence error). While the implementation for estimating divergence times is slightly different (e.g., our ABC approach uses time intervals between divergence events rather than absolute divergence times from the present), the results are very similar: very short internal branch lengths among gibbon genera and a total gibbon genera divergence time of ~5-6 Mya. However, as expected when using the full data rather than an approximation of it (Csilléry et al. 2010), the 95% CI estimated by G-PhoCS (Table S1) were substantially narrower than those estimated by ABC which are likely inflated because of a loss of information through the use of insufficient summary statistics.

**Allele Sharing and D-statistic analysis**

Because of the small sample sizes and large divergence times it is not expected that we would have the ability to infer gene flow if added as an additional parameter (whether an instantaneous pulse or continuous migration after divergence) in our ABC analysis. Although inter-genera hybrids have been observed in captivity, they are almost certainly infertile as a result of the complicated patterns of homology that would disrupt meiotic pairing. Moreover, such matings
have never been observed in the wild, even for sympatric species (Hirai et al. 2007). Therefore, it is unlikely that gene flow would continue for long after divergence as is typically modeled using isolation with migration approaches. Of course, this assumption depends on the rate of karyotypic change, which is thought to have occurred relatively soon after divergence and to have contributed to the speciation process (Carbone et al. 2014). Thus, accounting for biologically meaningful gene flow would increase the complexity of the model beyond what can likely be reliably inferred using ABC for this data set.

However, a fairly simply measure that can help to infer admixture events (although not necessarily help to reveal the mode, timing or extent of admixture) is the D-statistic (Durand et al. 2011). We first examined patterns of allele sharing across the whole genome by tallying the state of each genus at variable sites by a) choosing sites that met certain quality criteria (as determined by our masks) and that were homozygous for the same allele in both individuals from a genus (filt1) b), randomly sampling one allele from the two genotypes from a genus for sites that met the same quality criteria as a) (filt2), or c) randomly sampling one read from both individuals in a genus at a site (filt3) (Table S7a). We also repeated this at the species level, using only the highest coverage sample from each species (in this case filt1 reflects homozygous allele sharing) (Table S7c). Results were not qualitatively different using these different filtering criteria.

Consistent with our ABC analysis and Wall et al. (Wall et al. 2013), SSY and HLE share the largest number of alleles. Interestingly, while NLE and the two Hyllobates samples share a fairly low number of alleles compared with other pairwise comparisons, they both share more alleles with SSY than HLE. We performed a D-statistic analysis that demonstrated this excess sharing was statistically significant (Table S7b and d). Under the assumption that SSY
and HLE diverged last among the four genera as indicated in our ABC analysis, such a pattern is consistent with a model involving two independent gene flow events into SSY from both NLE and *Hylobates* after they diverged from HLE. An alternative model that does not invoke post-divergence gene flow involves the maintenance of long-term population structure between the ancestors of HLE and the ancestral population giving rise to the other gibbon genera (*Fig 2*). We attempted to incorporate population structure into our ABC framework but found via simulations that we could not distinguish between these models, especially given a parameter space consisting of short internal branch lengths as observed in this data set (data not shown).

We also used the D-statistic to examine whether there was any evidence of unbalanced allele sharing between the two *Hylobates* species. While the D-statistic slightly favored more allele sharing between HMO and the other three genera, the values were generally quite low and the Z-scores were only greater than |2| under filtering scheme 2.

**Discussion**

Previous attempts to resolve the phylogenetic relationships among the four gibbon genera based on different genetic systems (karyotypes changes, mtDNA, the Y chromosome, and short autosomal sequences, and ALU repeats) resulted in widely discordant phylogenies. All samples utilized in this study were also analyzed as part of the Gibbon Genome Project (*CARBONE et al. 2014*) where the best supported overall consensus tree based on genome-wide sequence divergence was found to be (((SSY, HLE), NLE),(HPI,HMO)). However, all four gibbon genera demonstrated a narrow range for sequence divergence (1.08-1.12%; mean 1.10%). Here, we developed a potentially powerful species tree analysis framework for four taxa that made use of genome-wide 2nd generation sequencing data and took into account discordant gene trees, and
applied it to the problem of the phylogenetic relationships of the four gibbon genera. Despite the availability of whole genome sequence data and the methodology demonstrating success with most simulated pseudo observed data sets, we could not confidently resolve the phylogenetic relationships between *Nomascus*, *Symphalangus*, *Hylobates* and *Hoolock*, although *Symphalangus* and *Hoolock* may represent the most recently diverged genera. This latter result is consistent with the best consensus gene tree identified by Carbone et al. (*Carbone et al.* 2014) and Wall et al. (*Wall et al.* 2013).

The most well supported bifurcating phylogeny is characterized by long external branch lengths and very short internal branch lengths, pointing to a rapid radiation of the four gibbon genera from a large ancestral effective population of \( \sim 10^5 \) individuals. Indeed, when we included an additional model representing a 4-way hard polytomy in our ABC analyses we found substantial support for this scenario of instantaneous divergence over any of the individual bifurcating topologies (at least the level of resolution of branch lengths afforded by the data). This demographic scenario would explain previous observations of genome-wide ILS (*Wall et al.* 2013) and discordant phylogenies across smaller datasets. However, we note that an alternative explanation is that the ancestral gibbon population already exhibited structure prior to the divergence of the four gibbon genera.

It is possible that that such a stark restructuring of the gibbon population during this proposed radiation event was driven by some major climatic or geological shift. This is particularly likely as gibbons reside predominantly on the relatively shallow Sunda continental shelf of Southeast Asia. At various times sea level changes and volcanic activity significantly altered the amount of habitable land (i.e., above sea level) in this region. As gibbons live a highly arboreal lifestyle, any reduction or fragmentation of their native forest habitats could have led to
extreme genetic isolation between geographically dispersed populations. This, coupled with a rapid evolution of karyotype differences, could have driven the speciation process among these gibbon taxa.

**Uncertainty in timing of the gibbon radiation**

It is important to note that associating the timing of speciation with the geological or climatological record is complicated by uncertainty in how we calibrate our estimates of $\tau$ (i.e., our choice of mutation rate). A phylogenetic estimate of $\mu$ for great apes that is often used is $\sim 1 \times 10^{-9}$ per site per year, an estimate based on calibrating sequence divergence with the fossil record (Takahata and Satta 1997; Nachman and Crowell 2000). This would place the radiation of gibbon genera within the early Pliocene $\sim 5$ Mya. Interestingly, it has been proposed that the Sunda shelf was largely one land mass up to 5 Mya (Outlaw and Voelker 2008), after which sea levels began to rise until $\sim 3$ Mya (Cichon et al. 2004) leading to the fragmenting of the region. There is evidence for an increased rate of divergence in other plants and animals during this early Pliocene window (Gorog et al. 2004; Outlaw and Voelker 2008; Akula et al. 2010; López-Guillerмо et al. 2010) and thus, it is possible that gibbon divergence may have been driven by the same process.

On the other hand, a value of $\mu = 0.5 \times 10^{-9}$ per site per year has recently been estimated using direct observation of mutations in human trios and quartets (Roach et al. 2010; Kong et al. 2012). Scally and Durbin (Scally and Durbin 2012) attempted to reconcile the phylogenetic and direct pedigree estimates with the fossil record (which itself is used to calibrate the phylogenetic estimate) by invoking the hominid slowdown hypothesis. Under this hypothesis, the increased body size of great apes correlates with a decrease in generation time and a
reduction in the annual mutation rate after their divergence from Old World monkeys. Evidence for this comes from evolutionary comparisons of Great Apes to Old World monkeys (e.g., humans have a 30% slower evolutionary rate as compared to baboons (Kim et al. 2006)). While generally bigger than Old World monkeys, the largest gibbons, from the genus Symphalangus, are approximately half the size of the smallest great ape, Pan paniscus. Thus, given that gibbons have smaller body sizes (and shorter generation times) than other apes, it is not clear to what extent the hominid slowdown hypothesis would apply.

Decreasing the mutation rate would lead to a Late Miocene speciation time of up to ~10 Mya, thus encompassing previous estimates of divergence at ~6-8 Mya based on mtDNA (Chan et al. 2010; Van Ngoc et al. 2010; Matsudaira and Ishida 2010). However, fossil calibration-based estimates such as used in these studies are subject to their own biases (Lukoschek et al. 2012), while estimates of demography from a single locus (especially a non-recombining region of the genome, no matter how well resolved the gene tree) are subject to large evolutionary stochasticity (Rosenberg and Nordborg 2002). It is noteworthy that the Y chromosome estimate differs from the mtDNA estimate substantially (5 and 9 Mya respectively) despite application of the same calibration procedures (Chan et al. 2012).

Our results do appear to rule out the hypothesis of Chivers (Chivers 1977), which suggests a Late Pleistocene divergence of gibbon genera. Despite this, the constant formation and destruction of land bridges during the Pleistocene that drives the Pleistocene pump hypothesis (Gorog et al. 2004; Akula et al. 2010) may have contributed to divergence of the several species within each gibbon genus (for example the pileatus/moloch split we observe ~1.6 Mya). Though the exact numbers are the subject of some debate, it is generally accepted that there are at least 7, 6 and 2 different Hylobates, Nomascus and Hoolock species respectively.
(Van Ngoc et al. 2010; Mittermeier et al. 2013; Rowe and Myers). Movement during these periods likely explains the current distribution of Hylobates species both on the mainland and the islands of Sumatra, Borneo, and Java, especially when one considers that gibbons probably cannot swim. Today neighboring gibbon species are largely isolated from each other by rivers. Further whole genome sequencing of multiple individuals from additional species, along with the application of powerful genomic methods to infer gene flow or admixture between species, will provide invaluable information for inferring the relationships among gibbon species across Southeast Asia. In addition, while it is well recognized that land bridges certainly formed during the Pleistocene, there is still great uncertainty as to whether these would have involved forest canopy or more savannah-like vegetation (Bird et al. 2005). Analysis of patterns of historic gene flow among the tree dwelling gibbons may help shed light on this process. Recent work using small amounts of autosomal sequence data (~11 kb) has already found evidence of asymmetrical gene flow between Hylobates species currently located on different islands (Chan et al. 2013) while a basic D-statistic analysis in this paper also hinted at the possibility of introgression between genera after divergence.

**Challenges in the use of whole genome sequence data for estimating demographic parameters**

Despite the fact that we generated whole genome sequences, it is important to appreciate that the explicit ABC modeling performed here utilized only a small amount of the total available data. A Pairwise Sequentially Markovian Coalescent (PSMC) (Li and Durbin 2011) analysis presented in Carbone et al. (Carbone et al. 2014) takes a different approach to utilizing genome scale sequence data. By incorporating patterns of genetic diversity across individual genome
sequences, important insights can be gained into changing $N_e$. To summarize these findings in the context the ABC demographic analysis presented here, both the NLE and HMO populations show major fluctuations in effective population size during the timeframe after gibbon genera diverged, when Pleistocene geological and climate shifts were taking place.

Composite likelihood methods that evaluate the entire allele frequency spectrum (AFS) across many populations may also prove useful for inference in situations such as the one presented here (Gutenkunst et al. 2009). These methods are likely to be particularly effective for estimating additional parameters such as recent size changes and migration when many individuals are sampled from multiple species compared to our approach, which uses only a summary of the AFS, though demographic events on the order of millions of years ago may still not be tractable regardless of sample size (Robinson et al. 2014). These approaches can also take advantage of recent methods that allow relatively unbiased estimation of the AFS from even low coverage 2nd generation sequencing data (though this currently must be done for each population separately). There have also been advances in extending these AFS-based approaches to many (more than three) populations and testing alternative scenarios even when using non-nested models and composite rather than full likelihoods (Excoffier et al. 2013), though whether these can reliably be used for contrasting many different models simultaneously in a phylogenetic context has yet to be explored. One notable limitation of these approaches is that they assume independence of sites and ignore linkage, and thus inference of migration and admixture is potentially underpowered (Sousa and Hey 2013).

However, to fully exploit whole genome data for demographic inference using coalescent methods it will be vital to construct genetic maps in gibbons, preferably separately for each genus, such that recombination can be appropriately incorporated into any population genetic
analysis. In addition, despite applying a correction factor in our analysis, reference bias towards *Nomascus* genomes was evident in our data, and it is likely that even more reference bias exists than we actually observe due to the variable karyotypes across genera. It seems unlikely that further large-scale Sanger sequencing will be used to link up scaffolds or generate reference genomes for the other three non-*Nomascus* genera, while short-read Illumina data is not suited for this task. However, the application of new sequencing technologies with long reads such as the PacBio (English et al. 2012) and nanopore (Schneider and Dekker 2012) technologies may provide useful and relatively low cost alternatives to assemble more robust reference genomes. This should lead to more powerful demographic and evolutionary analyses of gibbons in the future.

**Using ABC in Phylogenetics**

There is currently one published generalized ABC phylogenetic approach (ST-ABC) (Fan and Kubatko 2011). This method relies on having accurately phased sequence as it treats frequencies of gene tree topologies across loci as the data rather than summary statistics, and has only been tested and applied to relatively small datasets. However, it has also been questioned whether ST-ABC can accurately approximate the posterior distribution, as it relies on expectations of the distribution of gene trees rather than random simulations that incorporate sampling variability (Buzbas 2012). Our ABC approach does not have these limitations. While it could not reliably infer a particular bifurcating gibbon genera topology with any confidence because of the extremely short internal branch lengths (as we showed with our additional analysis of an instantaneous speciation model, a bifurcating topology may not actually exist in
this case of gibbons), simulations suggest our ABC approach should allow us to infer the correct species topology for four taxa in most reasonable cases.

However, it is important to appreciate that the framework applied here is tailored for this particular dataset involving unphased genome-wide data from a few individuals per taxa that diverged within the last 10 million years or so. How it would scale up with regard to speed with increasing numbers of samples, and how much accuracy and precision would be lost with fewer loci requires further investigation. It is possible that adding variance in the number of shared sites across loci as a summary statistic may prove useful in this case. In addition, increasing the number of taxa considered (even by one) could prove problematic due to a rapid increase in the parameter space (i.e., a large increase in the number of possible topologies) and an increase in the numbers of summary statistics needed to capture the phylogenetic structure (i.e., the potential impact of the “curse of dimensionality”). Combining more efficient ways of traversing tree space (BRYANT et al. 2012) may help with regard to the former issue, while choosing a more efficient set of summary statistics (e.g., via PLS) may improve the latter; however, there are still likely to be limits to how well the data can be summarized in just a few summary statistics for large phylogenies. Another potential issue of our approach that would place limits on the possible time depth for the phylogeny considered is the assumption of an infinite sites mutation model. It would be trivial to incorporate more complex substitution models, although this would also increase the computational burden.

With these improvements in mind, the ABC-family of methods have the potential to provide a useful and flexible phylogenetic tool that balances the need to incorporate large genomic datasets while taking into account gene tree uncertainty and variation in a coalescent framework. Genomic data is being generated at a rapid pace for a diverse set of species and it
clear is that phylogenetic methods are required than can accommodate such data. ABC provides one approach to do this.

**Acknowledgments**

Support for this work was provided by the National Institutes of Health to J.D.W and M.F.H (R01_HG005226) and an ERC Starting Grant (260372) and MICINN (Spain) BFU2011-28549 to T.M.-B. We thank Ryan Sprissler and the University of Arizona Genetics Core for assistance with sequencing and Ryan Gutenkunst for computing resources.
References


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


GOROG A. J., SINAGA M. H., AL E., 2004 Vicariance or dispersal? Historical biogeography of three Sunda shelf murine rodents (Maxomys surifer, Leopoldamys sabanus and Maxomys whiteheadi). Biological Journal of the ….


LI H., DURBIRN R., 2011 Inference of human population history from individual whole-genome


OUTLAW D. C., VOELKER G., 2008 Pliocene climatic change in insular Southeast Asia as an


ROWE N., MYERS M. (Eds.), All the world's primates.


Tables

Table 1: Gibbon samples undergoing 2\textsuperscript{nd} generation sequencing.

<table>
<thead>
<tr>
<th>Chr Nb</th>
<th>Genus</th>
<th>Species</th>
<th>Common Name</th>
<th>Code</th>
<th>Sex</th>
<th>Origin</th>
<th>Mean Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td><em>Nomascus</em></td>
<td>leucogenys</td>
<td>Northern white-cheeked</td>
<td>NLE</td>
<td>M</td>
<td>parents WB</td>
<td>13.78</td>
</tr>
<tr>
<td>50</td>
<td><em>Symphalangus</em></td>
<td>syndactylus</td>
<td>Siamang</td>
<td>SSY</td>
<td>M</td>
<td>sire WB, dam CB</td>
<td>12.80</td>
</tr>
<tr>
<td>38</td>
<td><em>Hoolock</em></td>
<td>leuconedys</td>
<td>Eastern hoolock gibbon</td>
<td>HLE</td>
<td>M</td>
<td>WB</td>
<td>19.15</td>
</tr>
<tr>
<td>44</td>
<td><em>Hylobates</em></td>
<td>pileatus</td>
<td>Pileated gibbon</td>
<td>HPI</td>
<td>M</td>
<td>parents WB</td>
<td>14.33</td>
</tr>
<tr>
<td>44</td>
<td><em>Hylobates</em></td>
<td>moloch</td>
<td>Javan gibbon</td>
<td>HMO</td>
<td>F</td>
<td>sire WB, dam CB</td>
<td>12.96</td>
</tr>
</tbody>
</table>

WB = wild born, CB = born in captivity
Table 2: Posterior probabilities for the 15 possible 4-population topologies for non-genic and genic loci

<table>
<thead>
<tr>
<th>Topology</th>
<th>Non-Genic</th>
<th></th>
<th>Genic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corrected</td>
<td>Uncorrected</td>
<td>Corrected</td>
<td>Uncorrected</td>
</tr>
<tr>
<td>(((SSY,HLE)(HPI,HMO))NLE)</td>
<td>0.16</td>
<td>0.15</td>
<td>0.19</td>
<td>0.15</td>
</tr>
<tr>
<td>(((HPI,HMO)NLE)SSY)HLE)</td>
<td>0.19</td>
<td>0.14</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>(((SSY,HLE)(HPI,HMO))NLE)</td>
<td>0.14</td>
<td>0.23</td>
<td>0.13</td>
<td>0.19</td>
</tr>
<tr>
<td>(((HPI,HMO)NLE)HLE)SSY)</td>
<td>0.13</td>
<td>0.11</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>(((NLE,HLE)SSY)(HPI,HMO))</td>
<td>0.06</td>
<td>0.05</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>(((HPI,HMO)SSY)NLE)HLE)</td>
<td>0.07</td>
<td>0.06</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>(((HPI,HMO)SSY)HLE)NLE)</td>
<td>0.05</td>
<td>0.07</td>
<td>0.07</td>
<td>0.14</td>
</tr>
<tr>
<td>(((HPI,HMO)NLE)(SSY,HLE))</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>(((NLE,SSY)HLE)(HPI,HMO))</td>
<td>0.03</td>
<td>0.03</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>(((NLE,HLE)(HPI,HMO))SSY)</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>(((NLE,SSY)(HPI,HMO))HLE)</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>(((HPI,HMO)HLE)SSY)NLE)</td>
<td>0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>(((HPI,HMO)HLE)NLE)SSY)</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>(((HPI,HMO)SSY)(NLE,HLE))</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>(((HPI,HMO)HLE)(NLE,SSY))</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Bold type indicates the topology identified using sequence divergence in Carbone et al. (CARBONE et al. 2014).
Table 3: Posterior probabilities for the 15 possible 4-population topologies as well as an instantaneous 4-way hard polytomy for non-genic and genic loci

<table>
<thead>
<tr>
<th>Topology</th>
<th>Non-Genic</th>
<th></th>
<th></th>
<th></th>
<th>Genic</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corrected</td>
<td>Uncorrected</td>
<td>Corrected</td>
<td>Uncorrected</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instantity</td>
<td>0.874</td>
<td>0.859</td>
<td>0.902</td>
<td>0.899</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((SSY,HLE)NLE)(HPI,HMO))</td>
<td>0.024</td>
<td>0.023</td>
<td>0.018</td>
<td>0.017</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)NLE)SSY)HLE)</td>
<td>0.024</td>
<td>0.023</td>
<td>0.011</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((SSY,HLE)(HPI,HMO))NLE)</td>
<td>0.023</td>
<td>0.037</td>
<td>0.016</td>
<td>0.018</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)NLE)HLE)SSY)</td>
<td>0.015</td>
<td>0.016</td>
<td>0.007</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)SSY)NLE)HLE)</td>
<td>0.008</td>
<td>0.007</td>
<td>0.010</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)NLE)(SSY,HLE))</td>
<td>0.007</td>
<td>0.006</td>
<td>0.005</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((NLE,HLE)SSY)(HPI,HMO))</td>
<td>0.007</td>
<td>0.007</td>
<td>0.009</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)SSY)HLE)NLE)</td>
<td>0.005</td>
<td>0.008</td>
<td>0.007</td>
<td>0.013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((NLE,HLE)(HPI,HMO))SSY)</td>
<td>0.003</td>
<td>0.003</td>
<td>0.004</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((NLE,SSY)HLE)(HPI,HMO))</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)HLE)NLE)SSY)</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((NLE,SSY)(HPI,HMO))HLE)</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)HLE)SSY)NLE)</td>
<td>0.002</td>
<td>0.004</td>
<td>0.002</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)SSY)(NLE,HLE))</td>
<td>0.001</td>
<td>0.001</td>
<td>0.003</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(((HPI,HMO)HLE)(NLE,SSY))</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Posterior estimates for an instantaneous speciation model for gibbon genera using a flat prior for $\tau$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>HDPI 95% fit$^a$</th>
<th>Posterior Estimation$^b$</th>
<th>Mode</th>
<th>Median</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{NLE}$</td>
<td>0.930</td>
<td>1.71E-03</td>
<td>1.72E-03</td>
<td>1.07E-03</td>
<td>2.73E-03</td>
<td></td>
</tr>
<tr>
<td>$\theta_{SSY}$</td>
<td>0.936</td>
<td>9.25E-04</td>
<td>9.24E-04</td>
<td>5.97E-04</td>
<td>1.43E-03</td>
<td></td>
</tr>
<tr>
<td>$\theta_{HLE}$</td>
<td>0.937</td>
<td>4.17E-04</td>
<td>4.17E-04</td>
<td>2.63E-04</td>
<td>6.58E-04</td>
<td></td>
</tr>
<tr>
<td>$\theta_{HPi}$</td>
<td>0.968</td>
<td>2.24E-04</td>
<td>2.25E-04</td>
<td>1.30E-04</td>
<td>3.92E-04</td>
<td></td>
</tr>
<tr>
<td>$\theta_{HMD}$</td>
<td>0.974</td>
<td>8.29E-04</td>
<td>8.32E-04</td>
<td>4.13E-04</td>
<td>1.68E-03</td>
<td></td>
</tr>
<tr>
<td>$\theta_{T1}$</td>
<td>0.958</td>
<td>3.54E-03</td>
<td>3.80E-03</td>
<td>7.69E-04</td>
<td>1.90E-02</td>
<td></td>
</tr>
<tr>
<td>$\theta_{Tanc}$</td>
<td>0.964</td>
<td>3.97E-03</td>
<td>3.97E-03</td>
<td>3.23E-03</td>
<td>4.86E-03</td>
<td></td>
</tr>
<tr>
<td>$\tau_1$</td>
<td>0.905</td>
<td>1.05E-03</td>
<td>1.23E-03</td>
<td>5.01E-04</td>
<td>2.18E-03</td>
<td></td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>0.911</td>
<td>2.69E-03</td>
<td>2.59E-03</td>
<td>1.41E-03</td>
<td>3.63E-03</td>
<td></td>
</tr>
</tbody>
</table>

$^a$A metric demonstrating how often known simulated values ($n=1,000$) fell within the calculated 95% CI, which gives a guide to the reliability of these CI’s for real data. $^b$Calculated using 15PLS components, 1,000,000 simulations and retaining 1%. All priors ranged from 0.0001-0.03 when $\log_{10}$ scaled.
Figures

Figure 1: Parameter estimates for the instantaneous radiation (A) and bifurcating speciation (B) model for gibbon genera. \( \mu = 7.5 \times 10^{-9} \) /site/generation, 10 years per generation. \( \theta_{T2} \) and \( \theta_{T3} \) based \( N_e \) values not to scale.
Figure 2: Cartoon of proposed model of ancestral population structure amongst gibbon genera.