The Annals of Applied Statistics
2010, Vol. 4, No. 4, 1722–1748
DOI: 10.1214/10-AOAS349

REUSE, RECYCLE, REWEIGH: COMBATING INFLUENZA THROUGH EFFICIENT SEQUENTIAL BAYESIAN COMPUTATION FOR MASSIVE DATA[1]

BY JENNIFER A. TOM, JANET S. SINSHEIMER AND MARC A. SUCHARD

University of California

Massive datasets in the gigabyte and terabyte range combined with the availability of increasingly sophisticated statistical tools yield analyses at the boundary of what is computationally feasible. Compromising in the face of this computational burden by partitioning the dataset into more tractable sizes results in stratified analyses, removed from the context that justified the initial data collection. In a Bayesian framework, these stratified analyses generate intermediate realizations, often compared using point estimates that fail to account for the variability within and correlation between the distributions these realizations approximate. However, although the initial concession to stratify generally precludes the more sensible analysis using a single joint hierarchical model, we can circumvent this outcome and capitalize on the intermediate realizations by extending the dynamic iterative reweighting MCMC algorithm. In doing so, we reuse the available realizations by reweighting them with importance weights, recycling them into a now tractable joint hierarchical model. We apply this technique to intermediate realizations generated from stratified analyses of 687 influenza A genomes spanning 13 years allowing us to revisit hypotheses regarding the evolutionary history of influenza within a hierarchical statistical framework.

1. Introduction.

1.1. *Studying the evolution of influenza A.* Influenza A continues to evade eradication resulting in ongoing economic and human cost. Yearly epidemics are responsible for 36,000 deaths on average in the United States. Three times in the past century global pandemics, including the infamous Spanish influenza of 1918, resulted in catastrophic mortality rates [Salomon and Webster (2009)]. Influenza epidemiologists believe a future influenza pandemic is an imminent threat [Webster and Walker (2003)]. Nearly 400 documented transfers [Salomon and Webster (2009)] of the highly virulent and potentially pandemic [Fauci (2005)] H5N1 strain of avian flu to humans in addition to the recent development of H1N1 swine flu [Butler (2009)] bolster the threat. The increasingly relevant necessity of

Received December 2009; revised March 2010.
[1]Supported by NIH Grants GM008185, GM086887, and GM053275.
Key words and phrases. Gibbs variable selection, hierarchical Bayesian model, importance sampling, influenza A, Markov chain Monte Carlo, massive data.

preventing future influenza pandemics requires a clear understanding of the evolutionary mechanisms of influenza as it is the key to vaccine development [Ghedin et al. (2005)].

Influenza A is a negative single-stranded RNA virus composed of 8 segments that total approximately 13 kb in length and encode 11 proteins. The three largest segments encode polymerases PB1, PB1-F2, PB2 and PA all of which are involved in RNA transcription and replication. The next three segments code for the two surface glycoproteins haemagglutinin (HA) and neuraminidase (NA) as well as the nucleoprotein (NP). The two smallest segments encode the matrix proteins M1 and M2 and the nonstructural proteins NS1 and NS2 [Yewdell and Garcia-Sastre (2002), Nelson and Holmes (2007)]. Influenza A research typically focuses on the epitope-rich HA and NA segments because they exhibit strong evolutionary selective pressure due to their direct interaction with the host immune system and are the primary determinants of the antigenic variation of influenza [Ghedin et al. (2005)]. The 16 HA and 9 NA glycoproteins found in the avian reservoir, referred to as H1 to H16 and N1 to N9, respectively, characterize and name the subtypes of influenza A [Nelson and Holmes (2007)].

The evolutionary history of influenza A involves the interaction of a number of mechanisms including mutation and reassortment. Approximately one random sequence mutation every replication cycle combined with the selective pressure on the surface glycoproteins results in an accumulation of point mutations on the HA and NA segments termed antigenic drift. The influenza genome also evolves through reassortment in which two subtypes coinfect a single host cell and exchange segments. This exchange of genetic material can lead to an antigenic shift or the creation of a new, potentially lethal, subtype. Reassortment between a virus in the avian reservoir and human influenza A resulted in the subtypes responsible for the Asian and Hong Kong influenza pandemics [Clercq (2006)] and the current swine flu pandemic derives from a triple reassortment event [Smith et al. (2009)].

As critically important events in influenza evolution occur at the genome-level, complete genome analysis yields scientific insight that single segments cannot afford. For example, Holmes et al. (2005) clarify a perplexing question in the evolutionary dynamics of HA by considering the varying histories of each segment. Analyzing 156 complete H3N2 viruses over a five-year time span from 1999 to 2004, Holmes et al. (2005) discover that while the Fujian-variant HA segment has been co-circulating since at least 2000, the variant only rises to dominance in 2002 after other segments within the influenza genome reassort and provide a synergistic background. This important reassortment event is only understood by studying the influenza genome in its entirety.

A more recent study by Rambaut et al. (2008) emphasizes the importance of incorporating model parameter uncertainty in drawing conclusions about influenza evolution through a Bayesian analysis of a truly massive dataset. Rambaut et al. (2008) compile 687 H3N2 influenza A full genomes sampled from New York over

a twelve-year period. Rambaut et al. (2008) address a host of biologically and clinically relevant questions including: (1) Are reassortment events coincident with shifts in HA antigenicity? (2) Do certain segments maintain greater genetic diversity? (3) Are the genetic histories of certain segments correlated? However, due to a dearth of Bayesian massive data techniques, computational constraints force Rambaut et al. (2008) to partition the data by stratifying on segment, using the data inefficiently, and drawing ad hoc conclusions about potential correlation. This current study rectifies the stratified analyses by fully capitalizing on the hierarchical nature of the influenza data and making formal inference after modeling the complete data in a single Bayesian analysis.

1.2. *Statistical context*. Despite optimized algorithms for missing data integration [Suchard and Rambaut (2009)], phylogenetic analysis of DNA sequences is lengthy and computationally intensive. Massive data measure in the gigabyte to terabyte range [Cressie, Olsen and Cook (1997)] and are increasingly common [Lambert (2003), Allison et al. (2009)]. This pervasiveness is particularly poignant in Bayesian models with missing data and especially in Bayesian models for stochastic processes where the dimensionality of the missing data can far outweigh the observed data. Such is the case in the evolutionary reconstruction of Rambaut et al. (2008).

One strategy pertinent to massive data inference is stratification [Cressie, Olsen and Cook (1997), Kettenring (2009)], often undesirable because it comes shackled with the host of difficulties arising from subgroup analysis [Glymour et al. (1997), Lagakos (2006)]. This identifies the direction that Rambaut et al. (2008) originally follow as they treat each of the eight segments independently. Shared computer memory and communication latency between computers limit hopes for considering a proper hierarchical model across segments simultaneously through which to share information and learn about segment-to-segment correlation. Even on state-of-the-art equipment, simulating sufficient realizations from posterior distributions conditional only on the data from a single segment, or what we refer to as "stratified distributions," still compels one to devote weeks of computing time per segment. This huge computational investment raises a critical point regarding a massive dataset with hierarchical structure. Often researchers perform preliminary analysis stratified by the exchangeable identifiers in the data simply because the statistical tools and computational resources exist for the stratified analysis. In attempting to fit the full hierarchical analysis, the ability to reuse the results from these suboptimal analyses represents a major savings in terms of time and resources and may even be the only feasible option.

To this end, we examine the dynamic iteratively reweighting MCMC algorithm (DyIRMA) [Liang and Weiss (2007), Liang et al. (2009)]. DyIRMA is based on the meta-analysis technique of using summary statistics from independent studies to infer a single hierarchical model [Carlin (1992), Warn, Thompson and Spiegelhalter (2002)]. Instead of summary statistics, however, DyIRMA combines realizations from independent distributions using importance sampling and Markov chain

Monte Carlo (MCMC) in an iterative process. Importantly, we can adopt DyIRMA to reuse realizations from preliminary stratified analyses. This desire to not waste intermediate realizations from the stratified analyses is particularly relevant in our influenza example because the realizations themselves require massive computing resources to generate. We further extend the insight of Liang and Weiss (2007) who combine intermediate realizations from two uncorrelated distributions. Our extension is necessary to accommodate correlated sequence data sampled over a span of thirteen years and allows us to entertain a much richer collection of hierarchical models motivated by the science at hand.

Our scientific aim in this study is to create a joint hierarchical model that addresses the questions raised by Rambaut et al. (2008) regarding the evolutionary history of influenza A. To this end, the hierarchical model must account for an unknown correlation structure between segments and allow for a flexible time-course in the model response, for which we exploit Gibbs variable selection (GVS) [Kuo and Mallick (1998), Dellaportas, Forster and Ntzoufras (2002)] to estimate a nonparametric response. The influenza A example illustrates that DyIRMA is a particularly flexible and valuable approach that reuses realizations via reweighting from computationally expensive distributions in a hierarchical framework. This widely applicable technique can be used to jointly model other independently generated, but in truth correlated, massive datasets.

As a preview, the paper continues as follows: Section 2.1 describes the generation of realizations from the stratified analyses, Section 2.2 introduces the basic framework used to estimate genealogies. Section 3 relates the machinery necessary to combine these realizations to estimate the joint hierarchical model along with computational concerns. Section 4 reviews the hierarchical model proposed, prior distributions, MCMC sampling concerns, and various modeling extensions. Sections 5 and 6 present results and conclude with a discussion.

2. Genomic-scale phylogenetic models.

2.1. *Intermediate phylogenetic realizations.* Rambaut et al. (2008) compile aligned sequence data for coding regions of each of the eight segments of the influenza A genome from the Influenza Genome Sequencing Project NCBI database [Ghedin et al. (2005)]. These alignments contain all 687 H3N2 influenza A genomes available over the 12 influenza seasons between 1993 and 2005. Season 2002 yields no sequences as it was predominantly an H1N1 season. From these data, Rambaut et al. (2008) are most interested in estimating and formally comparing the times to most recent common ancestor (TMRCA) of all the sequences sampled within each season for each segment. TMRCA can be thought of as a measure of genetic diversity because evolutionarily distant present-day sequences converge to a genealogy with longer branch lengths and consequently a larger TMRCA. To keep computation manageable, Rambaut et al. (2008) are forced to partition these data into independent blocks by segment. As the virus evolves through time,

samples from different seasons are highly interrelated through their shared history. Standard phylogenetic software packages account for this correlation. On the other hand, joint models across segments are less developed [Suchard et al. (2003)]. Consequently, initial analyses consider the segments independently, clouding conclusions about segment–segment interactions important to influenza A evolution.

Rambaut's et al. (2008) analyses provide samples from the intermediate distributions of TMRCAs given sequences from each individual segment. We are interested in the interaction of the evolutionary dynamics of influenza A segments over time and we use as our starting point realizations from these analyses stratified on segment. We let T_{ij} be the TMRCA for segment i and season j and $\mathbf{D}_i$ the sequence data for segment i. From each of the stratified analyses, we tabulate samples $\{T_{ij}^{(m)}|\mathbf{D}_i\}$ or $\{T_{ij}^{(1)}, \ldots, T_{ij}^{(M)}|\mathbf{D}_i\}$ for all (ij) where $m = (1, \ldots, M)$ indexes the MCMC sample, M is the total number of MCMC samples, $i = (1, \ldots, I)$ indexes the segment, and $j = (1, \ldots, J)$ indexes the season. Next, let $\mathbf{T}^{(m)}$ be the matrix constructed from J columns $\mathbf{T}_{\cdot j}^{(m)}$ or I rows $\mathbf{T}_{i\cdot}^{(m)}$, where $\mathbf{T}_{\cdot j}^{(m)}$ is an $I \times 1$ vector with all samples of TMRCA at iteration m for season j and $\mathbf{T}_{i\cdot}^{(m)}$ is a $J \times 1$ vector for segment i with all samples of TMRCA at iteration m for all J seasons. These M matrices $\mathbf{T}^{(m)}$ are the intermediate samples from the stratified distributions provided by Rambaut et al. (2008) that we propose to recycle into a hierarchical model.

To describe the construction of this stratified distribution, we first introduce some nomenclature. In brief, let g be the genealogy composed of a bifurcating acyclic graph (commonly called a topology) that describes the relatedness of a set of sequences and a vector of edge weights for the edges in this topology. Edges reflect the passage of time between bifurcation events and are also called branch lengths. Estimates of c different TMRCA, $\mathbf{T}$, embed in g because $\mathbf{T} = f(g)$ where $f(\cdot)$ is a deterministic mapping : $g \to \Re_{\geq 0}^{c}$. Here, $\mathbf{T}$ represents a $c \times 1$ vector containing the TMRCAs of interest and we use this general vector as the starting point before building up to $\mathbf{T}_{\cdot j}$.

In order to describe $f(\cdot)$, a brief introduction to the coalescent process is warranted [Hudson (1991)]. In the isochronous case, there are N sequences sampled at the same time t_0 where $t_0 = 0$ represents the present-day. Formation of a genealogy begins by randomly selecting two lineages at time t_e, $e = (1, \ldots, E)$ where e indexes the coalescent event and $E = (N - 1)$ is the total number of coalescent events. Proceeding back in time, the inter-coalescent time between the eth and $(e - 1)$th event is $u_e = t_e - t_{e-1}$ where $(u_1, \ldots, u_E)$ are independent exponential random variables. Let S be the set of all N taxa; then the E independent intercoalescent intervals yield an estimate T_S, the TMRCA of set S. Since each inter-coalescent interval is distributed as an exponential, the summation is a convolution of exponential distributions [Hein, Schierup and Wiuf (2005)].

We can estimate T for any subset of taxa. Let $\tilde{S} \subseteq S$ be a subset of the taxa, F represent the first event in this subset which for the isochronous case occurs

at $F = 1$, and L the last so that $0 < F \leq L \leq E$. Then $T_{\tilde{S}}$ is calculated similarly by identifying $g_{\tilde{S}}$, the subtree formed within g from the tips at t_0 and proceeding back in time to the last coalescent event occurring at t_L for all taxa in $\tilde{S}$. We then calculate $T_{\tilde{S}} = f(g_{\tilde{S}}) = \sum_{e=F}^{L} u_e$. Finally, as a nomenclature device, let $\tilde{G}$ be the set of all times of coalescent events identified in $\tilde{S}$, $\tilde{G} = \{t_F, \ldots, t_L\}$, and let $\tilde{G}^c$ be the complement. Then the genealogy can be decomposed into two disjoint sets of coalescent times as $g = (\tilde{G}, \tilde{G}^c)$, a device that will prove useful in Section 3.1.

One particularly interesting subset $\tilde{S}_j$ arises in the case of heterochronous data, namely T_j or TMRCA for all sequences sampled in a given season j. Because we have influenza A sequences sampled over time, our data are commonly called "heterochronous." We know the exact date of sampling for the influenza A sequences and can extend the coalescent to reflect this additional information. Elaborating on the description of the coalescent process above, the heterochronous case has two events of interest, coalescent and sampling, both of which can occur at multiple times. If there are O sampling times, there are now a total of $(N + O - 2)$ intervals, so in the heterochronous case, $E = (N + O - 2)$. Let $(t_0, \ldots, t_E)$ represent the times of coalescent or sampling events where t_0 is the most recent chronological sample and as before, the inter-event intervals are $u_e = t_e - t_{e-1}$. For $\tilde{S}_j$ identify the earliest event F and the last event L and again we have $T_{\tilde{S}_j} = f(g_{\tilde{S}_j}) = \sum_{e=F}^{L} u_e$.

We can now fully describe the intermediate realizations available from Rambaut et al. (2008). Recall that our research goals require we extract estimates of T_{ij} for a specific influenza season j where $j = (1, \ldots, J)$ indexes the J distinct seasons from a g_i for segment $i = (1, \ldots, I)$. We accomplish this by allowing the subset $\tilde{S} = S_{ij}$ to be the set of all taxa for a given influenza season j. Then $T_{ij} = f(g_{ij})$ is the summary statistic of interest for segment i and season j. Finally, let $\mathbf{T}_{i\cdot} = (T_{i1}, \ldots, T_{iJ})$ contain all estimates of TMRCA for a given segment and let $\mathbf{T}_{\cdot j} = (T_{1j}, \ldots, T_{Ij})$ be all estimates of TMRCA for a given year. Refer to Figure 1 for a simple example.

2.2. *Estimating TMRCA.* Although we have outlined how to generate our summary statistics of interest given a genealogy, further description is necessary regarding sampling from the distribution of the unknown genealogy g_i conditional on the data $\mathbf{D}_i$ for segment i. This posterior distribution can be represented as

$$P(g_i|\mathbf{D}_i) \propto \int\int \mathcal{L}(\mathbf{D}_i|g_i, \mathbf{Q}_i) P(g_i|\mathbf{\Phi}_i) P(\mathbf{\Phi}_i) P(\mathbf{Q}_i)\, \mathrm{d}\mathbf{\Phi}_i\, \mathrm{d}\mathbf{Q}_i, \tag{1}$$

where $\mathcal{L}(\mathbf{D}_i|g_i, \mathbf{Q}_i)$ is the likelihood of the sequence data given the genealogy and other phylogenetic parameters $\mathbf{Q}_i$ that model sequence change over time and $P(g_i|\mathbf{\Phi}_i)$, $P(\mathbf{\Phi}_i)$, and $P(\mathbf{Q}_i)$ are the prior distributions for the genealogy and phylogenetic parameters. The above decomposition specifies a marginal prior distribution on g_i because $P(g_i) = \int P(g_i|\mathbf{\Phi}_i) P(\mathbf{\Phi}_i)\, \mathrm{d}\mathbf{\Phi}_i$. We take note of this marginal

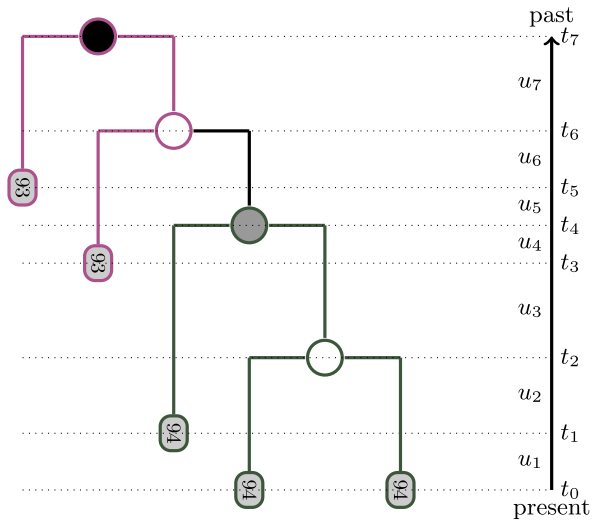

FIG. 1. Phylogenetic tree: calculating T, the time to most recent common ancestor (MRCA). *Five influenza A sequences are represented by gray rectangles and labeled with the sampling season. Present time is labeled as* t_0 *and extends back into the past until* t_7 *or the T for all five sequences. Inferred ancestral nodes between the samples are represented by circles. The black circle is the most recent common ancestor for both of the sequences sampled in the* 1993 *influenza season. We calculate* T_{1993} *by isolating the relevant branches, represented here by the purple subtree, and summing the inter-coalescent intervals,* $T_{1993} = u_4 + u_5 + u_6 + u_7$. *Similarly, the gray circle is the most recent common ancestor of all three sequences sampled in the* 1994 *influenza season and* $T_{1994} = u_1 + u_2 + u_3 + u_4$.

distribution because we wish to ultimately replace $\prod_i P(g_i)$ by a joint prior distribution $P(g_1, \ldots, g_I)$ in our hierarchical model. From Rambaut et al. (2008), $P(g_i)$ derives from a semiparametric relaxation of the coalescent process parameterized in terms of time-varying effective population size vector $\mathbf{\Phi}_i$ that follows a piecewise constant multiple-changepoint-process hyperprior distribution [Drummond et al. (2002, 2005)]. Effective population size is meant to reflect amount of genetic diversity rather than census count [Wright (1931)] and can be thought of as the average number of unique individuals that actually contribute genes to subsequent generations. Investigators generate samples from $P(g_i|\mathbf{D}_i)$ using MCMC in the Bayesian software BEAST [Drummond et al. (2002, 2005)] for each segment independently.

Due to the prominence the prior distribution $P(g_i)$ plays in the iterative reweighting scheme, it is outlined in some detail as follows, where we drop the subscript i for clarity. Recall that $(t_0, \ldots, t_E)$ are the times of events going into the past and $(u_1, \ldots, u_E)$ are the inter-coalescent intervals. Let $(k_1, \ldots, k_E)$ be the number of lineages that exist in g during a given inter-coalescent interval. We want to generate a sequence of effective population sizes of length B where $1 \leq B \leq E$

indexed by $b = (1, \ldots, B)$ with time similarly partitioned into $(\tilde{t}_1, \ldots, \tilde{t}_B)$. Essentially, we want to partition $\mathbf{\Phi}$ into B groups $(\phi_1, \ldots, \phi_B)$ where ϕ_b is constant between $\tilde{t}_b$ and $\tilde{t}_{b-1}$. In the heterochronous case, the number of lineages can increase (for a sampling event) or decrease (for a coalescent event) so there are two events of interest that can change the number of lineages. These events are distinguished by the indicator function $1_{\text{coa}}(e)$ which indicates that e is a coalescent event. Rambaut et al. (2008) specify the following heterochronous semiparametric coalescent prior distribution (again, ignoring dependence on i for clarity):

$$P(g|\mathbf{\Phi}) = \prod_{e=1}^{N+O-2} \left\{ \frac{k_e(k_e-1)}{2\phi_{h(e)}} \right\}^{1_{\text{coa}}(e)} \exp\left(-\frac{k_e(k_e-1)u_e}{2\phi_{h(e)}} \right), \tag{2}$$

where the function h maps from the larger number of E events to the B groups, or in other words $h(e) = b$ if t_e is between $\tilde{t}_b$ and $\tilde{t}_{b-1}$. Finally, to complete the specification of the prior distribution, the first effective population size follows a scale-invariant prior distribution, $P(\phi_1) \propto \frac{1}{\phi_1}$ [Drummond et al. (2005)] and the remaining $B-1$ effective population sizes are distributed as exponential with scale parameter equal to the previous effective population size, $\phi_b \sim \text{Exp}(\phi_{b-1})$. We make no claims about the appropriateness of this prior distribution choice. However, since the mean and variance of this prior distribution on ϕ_b grow with b and lead to some difficulty later, we point out that Minin, Bloomquist and Suchard (2008) provide a stable alternative with the joint skyride prior distribution on $P(g_i, \mathbf{\Phi}_i)$.

Now we have a foundation for understanding how genealogies are sampled and summarized using T_{ij} and what realizations from the intermediate distributions of the stratified analyses we have available. We wish to point out the benefits of the hierarchical model whose estimation we describe in the next section. These benefits include shrinkage estimators, a framework for statistical inference that accounts for correlation between strata, and models based on all available data. We now delve into how these independently generated estimates are combined into a joint statistical model reusing the preliminary realizations.

3. Computational recycling.

3.1. *Reweighting realizations.* In addition to the stratified realizations, the process of reweighting the stratified analysis samples of $\mathbf{T}_{i\cdot}$ to generate the joint hierarchical posterior distribution through DyIRMA requires $P(g_i)$ and the ability to evaluate the marginal prior distribution $P(\mathbf{T}_{i\cdot})$. We first present the DyIRMA machinery and then comment on how we extend it to accommodate this dataset. We save computational concerns regarding the calculation of $P(\mathbf{T}_{i\cdot})$ for Section 3.2.

Recall that we employ the following decomposition of genealogy $g_i = (\tilde{G}, \tilde{G}^c)$ which allows us to relate the summary statistic $\mathbf{T}_{i\cdot}$ to g_i as

$$P(\mathbf{T}_{i,\tilde{G}}|\mathbf{D}_i, \mathbf{\Omega}_0) = \int P(g_i|\mathbf{D}_i, \mathbf{\Omega}_0)\, \mathrm{d}\mathbf{T}_{i,\tilde{G}^c}. \tag{3}$$

The symbol $\boldsymbol{\Omega}_0$ identifies the stratified analyses and for notational consistency can be viewed as the forthcoming hierarchical model parameters, $\boldsymbol{\Omega}$ fixed at an arbitrary value. The subscript $\tilde{G}$ is dropped in the following equations to simplify notation.

We have intermediate realizations of the multivariate vector of $\mathbf{T}_{i\cdot} = (T_{i1}, \ldots, T_{iJ})$ under the individual models generated during the initial stratified analysis. We combine these realizations into a single joint posterior distribution conditional on all of the sequence data, $P(\mathbf{T}|\mathbf{D})$ where $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_I)$ and $\mathbf{T} = (\mathbf{T}_{1\cdot}, \ldots, \mathbf{T}_{I\cdot})$, by preferentially weighting samples that have a high likelihood under the hierarchical model relative to the probability of the prior distribution in the individual models. We make the following assumptions of conditional independence, namely that given $\mathbf{T}$ the sequence data are independent of the parameters in the hierarchical model, or $P(\mathbf{D}|\mathbf{T}, \boldsymbol{\Omega}) = P(\mathbf{D}|\mathbf{T})$. We also assume that given the hierarchical parameters, the TMRCAs $\mathbf{T}_{i\cdot}$, which we treat as exchangeable units, are independent or in other words we assume $P(\mathbf{T}|\boldsymbol{\Omega}) = \prod_i P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega})$. Given these assumptions, we can then use the following relationship:

$$P(\mathbf{T}|\mathbf{D}) \propto \int \prod_{i=1}^{I} \left(P(\mathbf{T}_{i\cdot}|\mathbf{D}_i, \boldsymbol{\Omega}_0) \frac{P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega})}{P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega}_0)} \right) P(\boldsymbol{\Omega}) \, \mathrm{d}\boldsymbol{\Omega}. \tag{4}$$

Therefore we reuse all available M realizations from $P(\mathbf{T}_{i\cdot}|\mathbf{D}_i, \boldsymbol{\Omega}_0)$ by identifying the following importance weights which are calculated for every sample of the stratified distributions:

$$w(\mathbf{T}_{i\cdot}, \boldsymbol{\Omega}) = \frac{P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega})}{P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega}_0)}. \tag{5}$$

The numerator of the weights, $P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega})$, is the conditional density under the hierarchical model. These hierarchical parameters are updated during each round of Gibbs sampling requiring recalculation of the weights for each iteration. The proposal density, $P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega}_0)$, is conditional on the parameters in the stratified analyses. As the variance of the estimate in (4) relies on the proposal density, some thought should be given toward selecting an appropriate prior distribution during individual analysis. Generally it is desirable to have a proposal density with heavier tails than the numerator of the weights [Robert and Casella (2004)].

As shown by Liang and Weiss (2007) and relying on the concept of importance sampling [Rubin (1988)], we can solve for the conditional distribution of the $\mathbf{T}_{i\cdot}$'s under the hierarchical model, which is specified as the following:

$$P(\mathbf{T}_{i\cdot}|\mathbf{D}, \boldsymbol{\Omega})_{\mathrm{DyIRMA}} = \frac{1}{W_i} \sum_{m=1}^{M} w\big(\mathbf{T}_{i\cdot}^{(m)}, \boldsymbol{\Omega}\big) \delta_{\mathbf{T}_{i\cdot}^{(m)}}(\mathbf{T}_{i\cdot}), \tag{6}$$

where $W_i = \sum_{m=1}^{M} w(\mathbf{T}_{i\cdot}^{(m)}, \boldsymbol{\Omega})$ and $\delta_{\mathbf{T}_{i\cdot}^{(m)}}(\mathbf{T}_{i\cdot})$ is a degenerate distribution at $\mathbf{T}_{i\cdot}^{(m)}$. The weighted stratified realizations are sampled during each round of Gibbs updates to generate samples from the hierarchical posterior distributions of $\mathbf{T}_{i\cdot}$ conditional on the current values of the parameters in the hierarchical model, $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(m)}$.

These samples in turn are then used to update the parameters in the hierarchical model from $\mathbf{\Omega}^{(m)}$ to $\mathbf{\Omega}^{(m+1)}$ which completes another iteration of sampling and contributes the last remaining piece of machinery necessary to jointly model our intermediate realizations from stratified analyses.

3.2. *Practical computational concerns*. We now describe in detail the strategy we employed to calculate $P(\mathbf{T}_{i\cdot})$ which is necessary for the calculation of weights. We include this description because for many scientifically interesting choices of the mapping $f(\cdot)$, such as jointly modeling TMRCAs, $P(\mathbf{T}_{i\cdot})$ remains intractable in analytic form. Fortunately, standard machinery already exists to draw simulants from this distribution, namely the MCMC sampler exploited to generate the stratified samples. A single additional run of this sampler without data provides all ingredients necessary to tackle the seemingly computationally intractable joint inference. Each partitioned dataset assumes identical prior distributions so a single simulation of $P(\mathbf{T}_{i\cdot})$ suffices for the reweighting of all the stratified distributions.

Two successful approaches to estimating marginal distributions from MCMC samples are multivariate kernel density [Scott (1992), Cacoullos (1964)] and importance weighted marginal density estimation (IWMDE) [Chen (1994)]. IWMDE proceeds by identifying a weighting function and sampling from a weighted ratio of the likelihood at a given value of the marginal of interest (here $T_{\tilde{G}} = t_{\tilde{G}}$) and realizations of the full joint distribution. The weighting function ideally has a similar shape to the unknown conditional marginal density, a Catch-22 circumvented with multivariate kernel density estimation. As a consequence of the strictly nonnegative support of $\mathbf{T}_{i\cdot}$ we control for potential boundary effects during kernel density estimation by using a gamma kernel [Chen (2000)] that removes the boundary effect and has the best mean integrated squared error among all nonnegative densities. For computational ease, we select the bandwidth using a multivariate adapted Scott's rule-of-thumb in which the bandwidth for the ath variable, $a = (1, \ldots, A)$, is $M^{1/(A+4)}\hat{\sigma}_a$ [Härdle et al. (2004)] where $\hat{\sigma}_a$ is the univariate sample standard deviation and bandwidths are allowed to vary for each univariate kernel resulting in a multiplicative kernel [Härdle (1990)].

The importance of selecting an appropriate prior distribution in a Bayesian framework is a topic of considerable depth [Gelman (2004), Efron (1986)]. Even the less contentious tactic of selecting a noninformative prior distribution by placing a uniform distribution over the parameter space can prove to be subjective [Kass and Wasserman (1996), Zwickl and Holder (2004)]. In the case where the likelihood function overwhelms the prior distribution, a potentially informative uniform prior distribution is rendered noninformative and specification of the prior distribution is deceptively unimportant. When Rambaut et al. (2008) specify a multiple changepoint prior distribution, the effective population sizes, $\mathbf{\Phi}_i$, were further constrained to lie between 0 and 120,000. This truncated prior distribution is well outside of the desired range of the values of $\mathbf{T}_{i\cdot}$, a fact that is inconsequential when sampling from the distribution conditional on the data. However, this creates

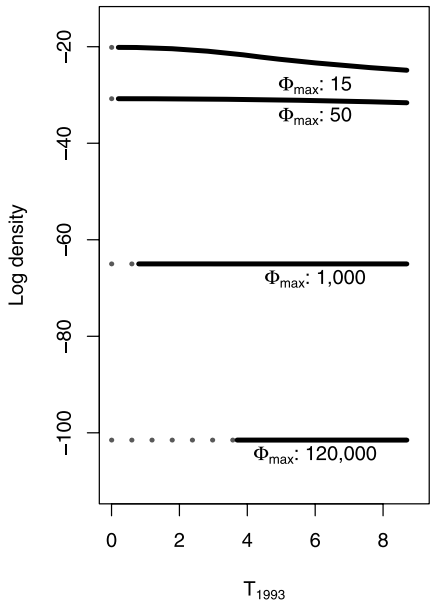

FIG. 2. *Prior density predicted by kernel density estimation* (*KDE*) *under different constraints on the prior distribution for a single representative season* (1993). *A prior density predicted by KDE with a gamma kernel conditional on the mean of the other eleven dimensions* $T_{1994}, \ldots, T_{2005}$. *We want the value of the prior density used during stratified analyses for the intermediate realizations sampled in the range illustrated in gray dot and black line. We have a KDE constructed with coverage from realizations of the prior distribution we generated in the range in black line. The gray dot region, or the region of interest the KDE is forced to extrapolate, expands as the maximum allowable* $\mathbf{\Phi}$, $\mathbf{\Phi}_{\max}$, *increases. The density with* $\mathbf{\Phi}_{\max} = 120{,}000$ *approaches the flat line illustrated with the perhaps more reasonable* $\mathbf{\Phi}_{\max} = 15$.

difficulties for evaluating the prior distribution at the realized values of the distribution conditional on the stratified data as they mostly lie well outside the region of the kernel density estimate. Luckily, when the maximum population size is constrained to be lower than 120,000 and coverage extends to the region of interest the KDE of the prior distribution reveals a relatively flat density surface.

As the maximum is gradually increased to 120,000, this surface decreases in the value of the density but remains relatively flat. This is illustrated in Figure 2, which shows a representative (T_{1993}) prior density predicted by KDE conditional on the mean of the other parameters for different maximum allowable $\mathbf{\Phi}_i$ sizes along with realizations from the prior distribution. The range of the values for the realizations from the prior distribution we wish to evaluate are indicated in the lower range of the density and the realizations used to predict the KDE are in the upper range. Taken together, this indicates the region the KDE is forced to extrapolate. As the addition of a constant value on the log scale does not affect the weights, a constant density equivalent to some arbitrary ε can be selected. For computational ease it

is desirable for ε to be similar in range to the density under the joint hierarchical model.

4. Hierarchical model: Antigenic shifts and diversity through time. Our methods enable us to conceive of a model that tests a greater range of hypotheses than those based on a single stratum. We revisit each of the statements of Rambaut et al. (2008) aimed at understanding the evolutionary history of influenza A with the advantages afforded by a hierarchical framework. We construct our model out of three basic elements. The first modeling element identifies seasons with a significant change in TMRCA from the previous year using GVS on the timepoints. A significant increase in TMRCA between timepoints suggests a reassortment event, whereas a significant decrease suggests a selective sweep. The second element introduces fixed segment effects that test whether certain segments have a higher TMRCA and therefore greater genetic diversity than others across time. Finally, we address correlation between the segments by exploring constrained variance matrices. To recycle random samples generated under existing stratified analyses, we implement an additional DyIRMA step during each round of Gibbs sampling of the joint model parameters. We first build up to the biologically motivated mean structure of the model containing GVS and fixed segment effects with an independent variance matrix and then introduce extensions for modeling the variance structure.

4.1. *Flexible modeling of time course through Gibbs variable selection.* At the most basic biological level, we must identify significant changes in TMRCA between influenza seasons and test if these correspond with shifts in HA antigenicity. We accomplish this with GVS as parameterized by Kuo and Mallick (1998) where the outcomes of the regression are the TMRCAs and season effects represent potential predictors. Note that this analysis would not be possible using a single segment as there would be insufficient degrees of freedom. The goal of Bayesian stochastic search variable selection is to identify the underlying generative model M_0 from the set of all possible models M. If J is the total number of possible predictors in the regression model, the model space has dimension 2^J, an arduous dimension from which to draw inference. Kuo and Mallick (1998) bypass this task by introducing indicator variables, $\gamma_j \sim \text{Bernoulli}(p_j)$, that identify the potential predictors of the outcome variable. To clarify, if the jth predictor has a marked effect on the outcome, the posterior probability that the corresponding indicator variable is one, $P(\gamma_j = 1|\mathbf{D})$, is high. On the other hand, if the jth predictor is not critical, the posterior probability that the corresponding indicator variable is zero, $P(\gamma_j = 0|\mathbf{D})$, is high. This implies that estimates of γ_j clarify which timepoints correspond to significant changes in TMRCA.

In order to proceed, we must introduce some additional nomenclature. Let $\mathbf{Z}$ be the $(IJ) \times (J-1)$ additive design matrix for all seasons where we remove the intercept in order to avoid overparameterization. In other words, $\mathbf{z}'_{ij}$

is a row vector where the first $(j-1)$ entries are 1 and the remaining $(J-j)$ are 0's. Then $\mathbf{Z}_j$ is made of I identical rows of $\mathbf{z}'_{ij}$ and stacking the J matrices of $(\mathbf{Z}_1, \ldots, \mathbf{Z}_J)$ generates $\mathbf{Z}$. Let $\boldsymbol{\Gamma}$ equal the diagonal matrix $\operatorname{diag}(\gamma_1, \ldots, \gamma_{J-1})$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{J-1})'$ be the unknown season effect sizes. The flexible GVS-induced mean time-course for MRCAs for all segments at season j becomes $\mathbf{Z}_j \boldsymbol{\Gamma} \boldsymbol{\beta}$ and is identical for all segments i. We select the following conjugate, independent, and noninformative priors distributions: $\beta_j \sim \mathrm{N}(\mu_\beta, \tau_\beta^{-1})$ and $\gamma_j \sim \mathrm{Bernoulli}(p_{j0})$ where μ_β, τ_β, and p_{j0} are hyperprior constants. Estimation of γ_j addresses questions about the evolution of influenza data over time and clarify whether shifts in HA antigenicity correlate with significant changes in TMRCA.

4.2. *Modeling the segment effect.* Modeling the segment effect highlights the importance of jointly modeling the influenza genome in concert in order to draw meaningful inference about similarities and differences in their evolutionary histories. Segment effects identify consistent differences in TMRCA over time and can test the hypothesis that NP has higher genetic diversity than HA. We also garner indirect information regarding the unresolved physical location of segments within the influenza A genome because we can resolve the correlation between segments and highly correlated segment histories are consistent with close proximity. Multidimensional scaling (MDS) suggests a relationship with decreasing intensity among the following three groupings: (1) {HA, M1/2}, (2) {NS1/2, NP}, and (3) {PA, PB1, PB2} [Rambaut et al. (2008)]. The NA segment is not grouped with any other segments. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)$ be unknown segment effects. Then we model

$$\mathbf{T}_{\cdot j} \sim \mathrm{N}(\boldsymbol{\theta} + \mathbf{Z}_j \boldsymbol{\phi}, \boldsymbol{\Sigma}), \tag{7}$$

where in its most general form $\boldsymbol{\Sigma}$ is assumed to be an $I \times I$ unstructured (UNS) covariance matrix. Conjugate prior distributions for this portion of the model are $\theta_i \sim \mathrm{N}(\mu_\theta, \tau_\theta^{-1})$ for all i and $\boldsymbol{\Sigma}^{-1} \sim \mathrm{Wishart}(\nu, \mathbf{R}^{-1})$ where μ_θ, τ_θ^{-1}, and ν are hyperprior constants and $\mathbf{R}^{-1}$ is the inverse of the hyperprior constant $(I \times I)$ scaling matrix. When combined with the prior distributions above, drawing realizations from the conditional posterior distributions for $\boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{\beta}$, and $\boldsymbol{\Gamma}$ helps address questions about segment effects, significant timepoints, and segment correlation simultaneously.

4.3. *Sampling from the complete model.* We now specify how to draw MCMC samples from the complete model. Let $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \boldsymbol{\beta})$ contain the unknown parameters of the hierarchical model, of which we wish to draw inference. Recall that we specify the conditional distribution for $\mathbf{T}_{i\cdot}$ in (6), which shows how we obtain realizations from $P(\mathbf{T}|\boldsymbol{\Omega}, \mathbf{D}) = \prod_{i=1}^{I} P(\mathbf{T}_{i\cdot}|\boldsymbol{\Omega}, \mathbf{D})$ by reweighting samples from $P(\mathbf{T}_{i\cdot}|\mathbf{D}, \boldsymbol{\Omega}_0)$. Therefore to estimate the parameters from this complete model using Gibbs sampling, we use DyIRMA during each Gibbs cycle over

$(\mathbf{T}, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \boldsymbol{\beta})$. We have stated the denominator of the weights previously as the predictions from the KDE of the stratified prior distributions that do not depend on $\boldsymbol{\Omega}$ and hence are constant during MCMC sampling. The numerator is updated at each iteration of MCMC and is simply the density of the vector of TMRCAs given the parameters in the hierarchical model, a straightforward way to determine multivariate normal. At this stage we are replacing the standard coalescent prior distribution with a normal prior distribution, a trade-off that yields straightforward interpretability as it allows us to directly test Rambaut et al. (2008) hypotheses, in addition to a simple computational implementation. For details on the other update steps for $(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\Gamma})$ and a schematic of sampling refer to the supplementary material [Tom, Sinsheimer and Suchard (2010)].

4.4. *Modeling extensions.*

4.4.1. *Constrained covariance matrices.* To identify segments with similar evolutionary histories, several constraints to $\boldsymbol{\Sigma}$ may provide more effective estimates. We explore an independent (IND) parameterization, such that $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_I$, with marginal variance σ^2 unknown that implies the evolutionary history of segments is not correlated. This specification allows inference to focus on the segment effects and has the additional advantage of substantially decreasing the number of inferred parameters. Also informative is compound symmetry (CS) which gives a general estimate of correlation between segments and provides a model nested within UNS to test for similar levels of correlation between segments. CS implies that the evolutionary histories of all segments are correlated with the same strength. Finally, autoregressive first order (AR1) and tridiagonal (TRI) structures with an estimable ordering of the segments directly identifies which segments have similar evolutionary histories. The motivation for nonexchangeable structures relies on the reasoning that segments with similar evolutionary histories have higher correlation than those with dissimilar histories.

For the CS model, we modify the Gibbs sampling by replacing the step for $\boldsymbol{\Sigma}$ with a Metropolis–Hastings step. Let $\boldsymbol{\Sigma} = \Psi(\boldsymbol{\xi}) = \Psi(\sigma^2, \rho)$ where ρ is the segment correlation and assume prior distributions $\sigma^2 \sim \text{Inverse-Gamma}(\alpha_{\sigma^2}, \lambda_{\sigma^2}^{-1})$ and $\rho \sim \text{Beta}(\alpha_\rho, \lambda_\rho^{-1})$ where $\alpha_{\sigma^2}, \lambda_{\sigma^2}^{-1}, \alpha_\rho$, and λ_ρ^{-1} are hyperprior constants. Refer to the supplementary material [Tom, Sinsheimer and Suchard (2010)] for further details on this modification to sampling.

4.4.2. *Finding the optimal correlation between segments.* Segments are not exchangeable in the TRI and AR1 parameterizations of the covariance matrix. In AR1, nearest-neighbor segments in the covariance matrix have higher correlation than those further apart. In TRI, the structure is more restrictive with segments more than one neighbor away from each other having no correlation. As the ordering of the segments i is not known and of paramount scientific interest, we estimate

the labeling or ordering of the segments within the covariance matrix by parameterizing a permutation vector $\boldsymbol{\pi}_v$ where $v = (1, \ldots, I!)$ indexes the different possible permutations. Sampling $\boldsymbol{\pi}_v$ requires an additional Metropolis–Hastings step. We propose $\boldsymbol{\pi}_v^*$ by randomly swapping two positions in the current permutation $\boldsymbol{\pi}_v$. We accept this proposal with probability $\min(1, \alpha_{\boldsymbol{\pi}_v})$ where $\alpha_{\boldsymbol{\pi}_v}$ is

$$\alpha_{\boldsymbol{\pi}_v} = \frac{P(\boldsymbol{\pi}_v^*)P(\mathbf{T}|\boldsymbol{\pi}_v^*, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \mathbf{D})}{P(\boldsymbol{\pi}_v)P(\mathbf{T}|\boldsymbol{\pi}_v, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \mathbf{D})}. \tag{8}$$

Assuming all permutations occur with equal probability, the terms $P(\boldsymbol{\pi}_v^*)$ and $P(\boldsymbol{\pi}_v)$ cancel. We proceed with caution when drawing inference due to label switching [Celeux, Hurn and Robert (2000)]. Instead of focusing inference on the permutation itself, we concentrate on the posterior probability that a given pair of segments are nearest neighbors, suggesting these segments have similar evolutionary histories. Our reasoning is that because only nearest neighbors are correlated, segments with similar evolutionary histories have a high posterior probability of adjacent positions within the permutation vector conditional on the correlation ρ being positive. This model that includes estimation of $\boldsymbol{\pi}_v$ (TRI-P) clarifies which evolutionary histories of segments are strongly correlated, which in turn has implications regarding structural interaction.

5. Results. Rambaut et al. (2008) have run stratified analyses for each of the eight influenza segments. Each analysis required an exhausting 2–3 weeks on high-end computers to approximate the stratified distributions $P(g_i|\mathbf{D}_i)$ via MCMC. Each MCMC chain runs for 10^8 iterations and subsampling every 10^5 iterations yields 10^3 approximately independent samples from these stratified distributions. We recycle these precomputed random samples to fit our hierarchical model that corrects for the stratification using the Bayesian machinery described in Sections 2 and 3. In particular, we implemented the DyIRMA Gibbs sampling scheme in cross-platform Java. We simulate three independent MCMC chains for each hierarchical model for 10^6 iterations with 10% burn-in and a 10-fold thinning. Each chain takes only approximately five hours to run on a mid-end desktop computer, representing a compelling and efficient alternative to fitting a joint hierarchical model starting from the sequence data. We assess the combined chains from three independent chains via several convergence criteria including trace plots, histograms, Geweke's convergence diagnostic [Geweke (1992)], and Rhat [Gelman (2004)].

Table 1 presents mean time-course estimates for the across-segments independent (IND) and tridiagonal with permutation (TRI-P) models. An additional subscript of IND or TRI-P clarifies the model for the parameter estimate. The additive parameterization of the design matrix results in parameter estimates that reflect an increase or decrease from the previous season. Table 1 therefore reports both the relative posterior conditional mean of β_j, $E(\beta_j|\mathbf{D}, \gamma_j = 1)$, that reflects the relative change in TMRCA (from the previous season) and, what we term, the absolute

TABLE 1

Posterior estimates for parameters summarizing time-course via Gibbs variable selection. *We report both the relative,* $E(\beta_j|\mathbf{D}, \gamma_j = 1)$, *and absolute,* $E(\theta_\cdot|\mathbf{D}) + E(\sum_{j=1994}^{J} \beta_j|\mathbf{D}, \gamma_j = 1)$ *where* $\theta_\cdot = \frac{1}{I}\sum_{i=1}^{I} \theta_i$, *values of TMRCA. Additive change is captured by posterior mean values for* β_j *conditional on selection in a given iteration for both the independent* (*IND*) *and tridiagonal with permutations* (*TRI-P*) *models. The superscript* † *indicates influenza seasons during which there was an HA antigenic shift*

Year	$P(\gamma_j = 1\|\mathbf{D})$	Relative posterior mean TMRCA	Absolute posterior mean TMRCA
Model—Independent (IND)			
1994	0.143	0.166	2.24
1995	0.113	0.110	2.35
1996	0.00667	0.00327	2.35
1997†	0.00115	0.0104	2.34
1998†	0.0748	−0.0665	2.30
1999	0.0207	−0.0128	2.28
2000	0.0133	−0.00605	2.27
2002	0.207	−0.242	2.03
2003	1.00	3.02	5.05
2004†	0.00852	0.00538	5.05
2005†	1.00	−3.86	1.20
Model—Tridiagonal (TRI-P)			
1994	0.0500	0.0568	2.11
1995	0.0356	0.0326	2.14
1996	0.00593	0.00235	2.14
1997†	0.00556	0.00122	2.14
1998†	0.00100	−0.00511	2.14
1999	0.00815	−0.00346	2.13
2000	0.00852	−0.00268	2.13
2002	0.0437	−0.0456	2.09
2003	1.00	2.89	4.96
2004†	0.0159	0.0118	4.98
2005†	1.00	−3.79	1.18

posterior conditional mean or posterior conditional mean of the segment effect average added to the cumulative posterior mean of β_j, $E(\theta_\cdot + \sum_{j=1994}^{J} \beta_j|\mathbf{D}, \gamma_j = 1)$ where $\theta_\cdot = \frac{1}{I}\sum_{i=1}^{I} \theta_i$. Additionally, we report the posterior mean estimates for γ_j that reflect the posterior probability that β_j is included in the model.

Looking first at the IND model, an indicator estimate of 1 at season 2003 suggests that there exists a significant difference between the average segment time to MRCA, T, at influenza season 2002 and that of 2003 because the posterior probability that the inclusion of a regression parameter captures this difference approaches 1. This decisive support of a significant jump in T suggests an infusion of genetic diversity and is consistent with a *reassortment event*. Similarly, the

posterior probability for the indicator representing the shift between seasons 2004 and 2005 also approaches 1. However, in this instance, as the negative sign of the estimate for $\beta_{\text{IND},2005}$ reflects, this suggests a decrease in genetic diversity which is consistent with a *selective sweep*.

Looking more closely at the IND model, the 2003 shift corresponds to an increase in T with a posterior mean regression parameter estimate of $\hat{\beta}_{\text{IND},2003} = 3.02$ and 95% Bayesian credible interval (BCI) $(2.23, 4.28)$. The 2005 shift suggests a $\hat{\beta}_{\text{IND},2005} = -3.86$ $(-4.78, -3.00)$ decrease in T and is concomitant with the FU02-CA04 HA antigenic shift. No other indicators have posterior probabilities greater than 0.95. Similarly, the TRI-P model furnishes strong support for including seasons 2003 ($\hat{\gamma}_{\text{TRI-P},2003} = 1$) and 2005 ($\hat{\gamma}_{\text{TRI-P},2005} = 1$) and shows the similar pattern of jump in T of $\hat{\beta}_{\text{TRI-P},2003} = 2.89$ $(2.00, 3.91)$ followed by decrease of $\hat{\beta}_{\text{TRI-P},2005} = -3.79$ $(-5.03, -2.55)$. The increase in genetic diversity at season 2003 followed by a decrease at 2005 clearly identify themselves in the third column of Table 1 that shows that the posterior conditional segment means for the seasons between 1994 and 2002 as around 2 but jumps to nearly 5 in seasons 2003 and 2004 before undergoing a decline in 2005 and decreasing to around 1.2. Figure 3 reiterates this finding. In general, the TRI-P indicator probabilities are closer to 0 or 1 than in the IND model. For example, the posterior probability of the season 2002 shift ($\hat{\gamma}_{\text{TRI-P},2002} = 0.0437$) is less than that for the IND model ($\hat{\gamma}_{\text{IND},2002} = 0.207$). One explanation for the differences between the two models is that the parameter estimate standard errors are generally reduced for the TRI-P model.

There exists a wide number of covariance structures that are biologically interesting to explore within our framework. Specifically, we also consider the unstructured (UNS) and first order autoregressive with permutation (AR1-P) models. Although we observe a similar pattern of increase and decrease in TMRCA in 2003 and 2005, these models that allow a high level of correlation between the segments lose identifiability of the indicator variables across time, presumably because the variation present in the observations is used to model the correlation instead of the mean structure. Combined with the fact that inference focuses on the mean structure, we continue to exclusively discuss results for sparse covariance matrices such as those used in the IND and TRI-P models.

Summarizing the posterior probability of a model that contains all of the antigenic shifts addresses our question regarding whether HA shifts are concomitant with significant changes in T. Antigenic shifts occur at 1997 (BE93-WU95), 1998 (WU95-SY97), 2004 (SY97-FU02), and 2005 (FU02-CA04) and these seasons are superscripted in Table 1. We refer to these seasons from now on as shifts 1, 2, 3, and 4, respectively, and use these shift numbers as subscripts on a given model M to clarify which model variables are included. For example, M_{13} describes a model where both the first and third antigenic shifts are included ($\hat{\gamma}_{1997} = 1$ and $\hat{\gamma}_{2004} = 1$). As we have already noted, the posterior probability that the FU02-CA04 shift in 2005 is included in both the IND and TRI-P models approaches 1.

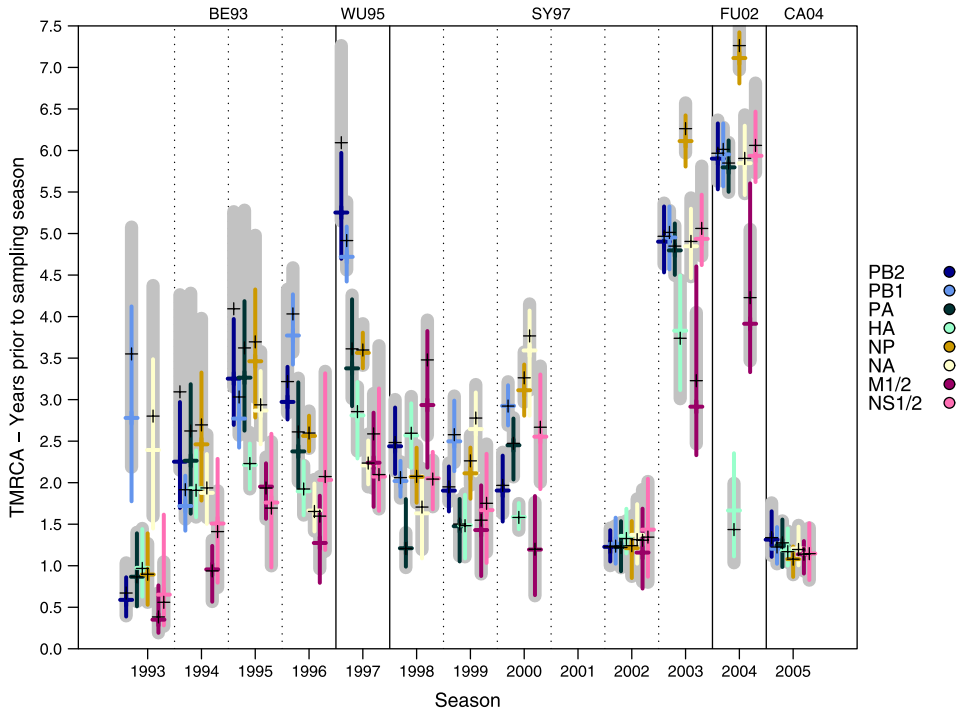

FIG. 3. Realizations from stratified and hierarchical analyses. *Stratified* (*gray*) *and hierarchical* (*colored line*) *sample means and* 95% *highest density intervals conditioned on data for the eight segments* (*PB*2, *PB*1, *PA*, *HA*, *NP*, *NA*, *M*1/2, *and NS*1/2) *of H*3*N*2 *influenza A listed from largest to smallest. Samples from the hierarchical posterior distributions are from the tridiagonal with permutations* (*TRI-P*) *model. Twelve seasons are depicted with season* 2001 *missing due to an H*1*N*1 *dominant season. Segment distributions are staggered for clarity. The y-axis represents TMRCA and HA antigenicity* (*BE*93, *WU*95, *SY*97, *FU*02, *and CA*04) *is indicated across the top. Hierarchical posterior distributions exhibit shrinkage toward the mean relative to the stratified.*

However, the posterior probability of this being the only antigenic shift selected is also large because $P_{\text{IND}}(M_4|\mathbf{D}) = 0.91$ and $P_{\text{TRI-P}}(M_4|\mathbf{D}) = 0.97$. The posterior probability of a model with all four antigenic shifts, $P(M_{1234}|\mathbf{D})$, approaches zero for both TRI-P and IND. This strongly suggests that HA antigenic shifts are not strictly concomitant with significant changes in TMRCA from the previous season after correcting for the correlation structure of the segments. Rambaut et al. (2008) are unable to correct for the correlation structure across segments. Given the limited time series data available relative to the number of antigenic shifts, more data would certainly enhance the understanding of the relationship between reassortment events, selective sweeps, and HA antigenic shifts.

To more thoroughly unravel the evolutionary history of influenza A, we need to address the relative level of evolutionary diversity between segments. To this end we include segment-specific effects whose posterior estimates we summa-

TABLE 2

Posterior mean and 95% Bayesian credible intervals (BCIs) for segment-specific effects. *Posterior mean and probability intervals for segment effects in both the independent (IND) and tridiagonal with permutations (TRI-P) models. The superscript † indicates our comparison of particular interest. The posterior probability of NP being greater than HA is* 0.995 *in the IND model* ($BF_{\text{IND}} = 179$), *assuming equal prior probability. Similarly, for the TRI-P model, the posterior probability is* 0.941 ($BF_{\text{TRI-P}} = 16.1$). *Both provide strong evidence that the evolutionary history of NP has greater diversity than that of HA*

	Posterior mean	**95% BCI**
Model—Independent		
PB2	2.51	(1.30, 3.40)
PB1	2.65	(1.42, 3.49)
PA	2.11	(0.888, 2.96)
HA†	1.37	(0.161, 2.17)
NP†	2.53	(1.30, 3.36)
NA	2.21	(1.01, 3.02)
M1/2	1.39	(0.244, 2.19)
NS1/2	1.78	(0.585, 2.62)
Model—Tridiagonal		
PB2	2.35	(1.24, 3.17)
PB1	2.52	(1.29, 3.37)
PA	2.08	(1.03, 2.91)
HA†	1.54	(0.381, 2.47)
NP†	2.46	(1.25, 3.28)
NA	2.20	(1.04, 3.04)
M1/2	1.37	(0.168, 2.44)
NS1/2	1.88	(0.770, 2.79)

rize in Table 2 for both the IND and the TRI-P models. We report both the posterior mean and the 95% BCIs with the range of the TRI-P intervals generally slightly reduced from the IND, perhaps indicative of the TRI-P being a more appropriate covariance structure with which to model the correlated parameters of the different segments. PB1 returns the highest posterior mean estimates of T which at season 1993 for IND is $\hat{\theta}_{\text{IND,PB2}} = 2.51$ (1.30, 3.40) and for TRI-P is $\hat{\theta}_{\text{TRI-P,PB2}} = 2.35$ (1.24, 3.17). HA and M1/2 yield the lowest posterior means of T which, also given at season 1993, are $\hat{\theta}_{\text{IND,HA}} = 1.37$ (0.161, 2.17) and $\hat{\theta}_{\text{IND,M1/2}} = 1.39$ (0.244, 2.19) respectively for IND and $\hat{\theta}_{\text{TRI-P,HA}} = 1.54$ (0.381, 2.47) and $\hat{\theta}_{\text{TRI-P,M1/2}} = 1.37$ (0.168, 2.44) respectively for TRI-P. This implies that genetic diversity is maintained longer in PB1 than HA and M1/2 with the difference between these segments for T on the order of an entire year. We go further than Rambaut et al. (2008) who simply observe the differences in the T of different segments of influenza A by formally testing whether certain segments

maintain greater diversity than others. One relationship of particular interest, explored but not formally tested by Rambaut et al. (2008), is the comparison between NP and HA, superscripted in Table 2. When assuming equal prior probability of both outcomes (0.5 each), the Bayes factors (BF) of whether the T of NP is greater than HA are 179 for the IND model and 16.1 for TRI-P. This means that the posterior probability of NP being greater than HA in the IND model is 0.994 and 0.941 in the TRI-P, providing strong support for the hypothesis that NP maintains greater genetic diversity than HA.

Finally, we approach teasing out the correlation between segments by estimating the posterior distribution of all possible segment order permutation within the tridiagonal covariance matrix in the TRI-P model. The three groupings of {HA, M1/2}, {NP, NS1/2}, and {PA, PB1, PB2} originally posited by Rambaut et al. (2008) arise from ad hoc multidimensional scaling results. For notational convenience, we now refer to these segments by number and place them in alphabetical order as 1 (HA), 2 (M1/2), 3 (NA), 4 (NS1/2), 5 (NP), 6 (PA), 7 (PB1), and 8 (PB2). This notational device means that we refer to a model with both the {HA, M1/2} and {PA, PB1, PB2} grouping as $M_{\{12\}\{678\}}$. Note that there is no implied ordering of the remaining unlisted segments so that $M_{\{12\}}$ and $M_{\{678\}}$, say, have some overlapping groupings. There are seven pairs of neighbors possible in the tridiagonal matrix because the two segments occupying the corners of the diagonal of the covariance matrix are restricted to having a single neighbor and the middle six segments can each have two neighbors.

We are interested specifically in positive correlation between segments as we want to clarify which segments have similar evolutionary histories. With this purview, we focus on results conditional on the correlation ρ being greater than 0.2; this provision occurs with approximately 0.593 posterior probability, but eases interpretation. In general, we assume there are two segments i and i' where $i, i' \in \{1, \ldots, 8\}$, and $i \neq i'$ which are selected as neighbors. We summarize the posterior probability of models in which two segments are neighbors for our subset of interest, $P(M_{\{ii'\}}|\mathbf{D}, \rho > 0.2)$, in Table 3. The posterior probability of the strongly hypothesized {HA, M1/2} group is very high, $P(M_{\{12\}}|\mathbf{D}, \rho > 0.2) = 0.768$. The posterior probability of the {PB1, PB2} pairing is also high, $P(M_{\{78\}}|\mathbf{D}, \rho > 0.2) = 0.493$. Previously unidentified as a potential pairing is {NP, PB2}, $P(M_{\{48\}}|\mathbf{D}, \rho > 0.2) = 0.378$ as well as {NP, PA}, $P(M_{\{46\}}|\mathbf{D}, \rho > 0.2) = 0.313$ which implies that the NP segment might be just as strongly aligned with the {PA, PB1, PB2} grouping as with NS1/2 because the posterior probability of the {NP, NS1/2} grouping is similar, $P(M_{\{45\}}|\mathbf{D}, \rho > 0.2) = 0.335$.

Again, we formally test these hypothetical groupings against the null hypothesis that all permutations of the tridiagonal covariance matrix are equally likely and summarize the results in Table 4. We make these calculations unconditional on the correlation ρ. Notice in column 2 of Table 4, the posterior probability of grouping {HA, M1/2} is 0.624 and has a prior probability of 0.0357 leading to a significant BF of 17.5. Again, {NP, NS1/2} has a weaker BF of 8.50 lending some doubt to

TABLE 3

Posterior probability of segments as neighbors. *The values represent the posterior probability of two segments having correlation in the tridiagonal with permutation (TRI-P) model conditional on the correlation being greater than* 0.2 *which occurs with* 0.593 *probability. For notational convenience we refer to the segments by number*, 1 (*HA*), 2 ($M1/2$), 3 (*NA*), 4 (*NP*), 5 (*NS*), 6 (*PA*), 7 (*PB*1), *and* 8 (*PB*2), *so the model with HA and* $M1/2$ *as neighbors is* $M_{\{12\}}$. *Hypothesized structural groupings are indicated by the superscript*, $\{HA, M1/2\}$ *is* $\{12\}$, $\{NS1/2, NP\}$ *is* $\{45\}$, *and* $\{PA, PB1, PB2\}$ *is a subset of* $\{678\}$. *The posterior probability of* $\{HA, M1/2\}$ *being grouped as neighbors is the highest* ($P(M_{\{12\}}|\mathbf{D}, \rho > 0.2) = 0.768$) *suggesting these two segments share similar evolutionary histories*

	PB2	**PB1**	**PA**	**HA**	**NP**	**NA**	**M1/2**	**NS1/2**
PB2	—	$0.493^{\{78\}}$	$0.297^{\{68\}}$	0.181	0.378	0.123	0.185	0.165
PB1		—	$0.284^{\{67\}}$	0.151	0.257	0.313	0.0990	0.160
PA			—	0.192	0.313	0.230	0.173	0.290
HA				—	0.0620	0.154	$0.768^{\{12\}}$	0.157
NP					—	0.345	0.0880	$0.335^{\{45\}}$
NA						—	0.222	0.283
M1/2							—	0.304
NS1/2								—

the hypothesis that these two segments have similar evolutionary histories. Finally, the posterior probability that all three pairings are found is 0.0741 which, given all models are thought a priori to have equal probability, has a prior probability of 0.000600. This leads to a very decisive BF of 124 which strongly supports Rambaut's et al. (2008) hypothesized groupings of segments.

We have already demonstrated the ease with which BF can be assigned to competing joint hierarchical analysis models once we have recycled the realizations from stratified analyses into a single statistical framework. However, an additional

TABLE 4

Bayes factors (BFs) of hypothesized structural groupings being nearest neighbors. *BFs are calculated for the tridiagonal with permutations (TRI-P) model testing the support for the hypothesized structural groupings of* $\{HA, M1/2\}$, $\{NS1/2, NP\}$, *and* $\{PA, PB1, PB2\}$. *The posterior odds, prior odds, and BFs are reported for each grouping individually and then for all three occurring together. All possibilities of groupings are considered equally probable for the prior odds. There is strong support for the* $\{HA, M1/2\}$ *pairing* ($BF = 17.5$) *and decisive support for all three structural groupings being selected as neighbors* ($BF = 124$)

	{HA, M1/2}	**{NP, NS1/2}**	**{PA, PB1, PB2}**	**All three**
Posterior odds	0.624	0.304	0.195	0.0741
Prior odds	0.0357	0.0357	0.0179	0.000600
Bayes factor	17.5	8.50	10.9	124

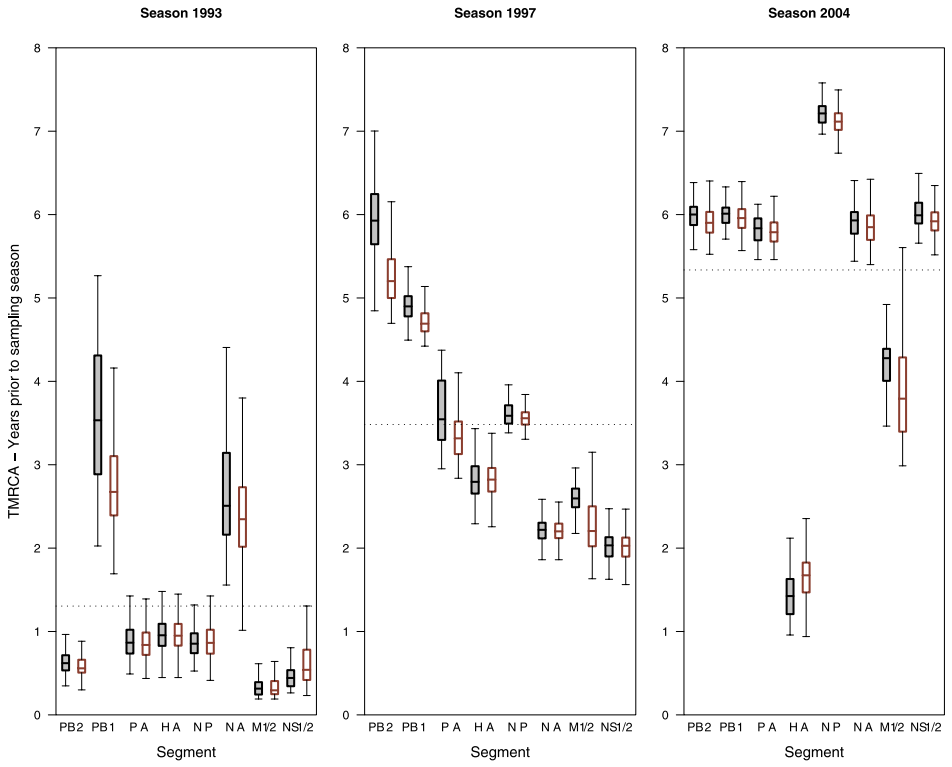

FIG. 4. Box plots of stratified and hierarchical estimates for select seasons. *Box plots of stratified* (*gray*) *and hierarchical* (*blue*) *for the selected seasons* 1993, 1997, *and* 2004. *On the x-axis are segments listed from left to right in descending order of size, and on the y-axis are adjusted TMRCA. Selected seasons clearly display shrinkage toward the mean of the hierarchical versus stratified distributions.*

advantage is that this hierarchical posterior distribution demonstrates shrinkage toward the mean when compared against the stratified results, leading to more sensible segment-specific estimates. Figure 3 displays the marginal hierarchical posterior distributions superimposed on the stratified results for all eight segments over time. In Figure 3 we demonstrate that overall the direction of shrinkage estimates from the stratified mean (black) to the hierarchical mean (color) draws toward the grand mean for that time. Shrinkage toward the mean especially tempers outliers, a phenomenon illustrated in Figure 4. Focusing on 1993, this season emits positive outliers in PB1 and NA. Season 1997 has positive outliers in PB2 and PB1 and 2004 yields a negative outlier in HA and a positive outlier in NP. In 1993, coverage drastically shifts down for PB1 and NA as well as shifts up for M1/2 and NS1/2. In season 1997, coverage greatly decreases and shifts downward for the outlier PB2 and in season 2004, coverage shifts up for HA and shifts down for NP. Figure 3 also illustrates further advantages of this technique, including the

reduced credible intervals for the hierarchical distribution relative to the stratified. Finally, especially apparent in this figure is a cautionary reminder of the wide variability present in the inappropriately independent, stratified results. For example in 1993, the stratified estimate uncertainty of PB1 spans around 3 years whereas that for PB2 spans around 0.5. These differences in variability mitigate when basing conclusions on the point estimates of the stratified results. Therefore the advantages of reusing the stratified analyses in a joint model lie not only in the ability to assign across-segment BFs and the incorporation of highly desirable shrinkage estimators that lead to improved estimation [Efron and Morris (1977)], but also in enhanced modeling capabilities that more accurately represent the variability in the segments.

6. Conclusions. Increasing dataset sizes are engulfing the scientific community [Anderson (2008)] demanding novel approaches to statistical analysis. While the introduction of GPU programming to the statistical community promises solutions in the near-future [Suchard et al. (2010)], the daunting task of analyzing these massive datasets is currently made realistic by partitioning them into smaller, more tractable sizes. This stratification, while facilitating fast estimation, results in overparameterization and ignores the correlation between parameters across strata. Additionally, stratification fails to profit from the massive amounts of data available because parameters are estimated from siloed strata, removed from the implicit context that motivated the initial data collection.

Ideally, given no computational constraints, related and exchangeable groups are represented by a hierarchical model. This framework efficiently pools information across groups while accounting for the correlation between them. This single unified model also makes it easy to draw dataset-wide inference. Finally, hierarchical models lead to improved estimators due to shrinkage toward the mean; this well-known phenomenon is termed Stein's paradox [Efron and Morris (1977)].

Perhaps a more familiar approach to constructing this full hierarchical model is sequential Monte Carlo (SMC) [Doucet, de Freitas and Gordon (2001)]. Chopin (2002), Ridgeway and Madigan (2003), and Balakrishnan and Madigan (2006) use particle filtering as a SMC solution to the massive data problem, building up the posterior distribution of the complete data by incrementally introducing a small number of data into the posterior distribution using importance sampling [Cappé, Godsill and Moulines (2007)]. This form of SMC is highly effective for linearly organized data such as time series observations but is inappropriate for data divided into exchangeable groups. Further, particle filters do not completely recycle the preliminary analyses.

Our methods in this paper create a new strategy, combining the advantages of stratification, namely speed, with the statistical framework of hierarchical modeling. Any hypothesis addressed in the subpar stratified model can be reused, benefiting from assigned measures of statistical certainty. Our methods capitalize on

the intermediate realizations from stratified analyses, recycling them into the hierarchical model by reweighting via importance sampling. From the standpoint of the evolutionary history of influenza A, Rambaut et al. (2008) are ambitious in their goal of understanding individual segments within the larger context of the complete genome. Our methods enable us to revisit Rambaut et al. (2008)'s conclusions with the insight afforded by a hierarchical statistical framework.

We find that for these data, the TRI-P model is quite sufficient for our reexamination of the biological questions. However, in some circumstances it may be necessary to allow for some small degree of correlation between segments that are not nearest neighbors. One way to accommodate this correlation is by assigning a prior inverse-Wishart distribution to $\mathbf{\Sigma}$. Our approach is to center the inverse-Wishart distribution on the structured covariance matrix [as in Boscardin and Zhang (2004)]. The degrees of freedom of the inverse-Wishart provide a tuning parameter. As the degrees of freedom go to infinity the extra correlation goes to zero and we recover our original model. The use of this model requires that we replace the Gibbs sampling step for $\mathbf{\Sigma}$ with a Metropolis–Hastings step, and so adds some computational burden.

The applications of reusing, recycling, and reweighting are limited only by the biological questions of interest. This flexible framework has far-reaching value into areas such as resequencing and phylogeography [Knowles (2004)], in other words, situations where computational complexity forces data partitioning, preventing the more appropriate hierarchical model. From an applied statistician's perspective, this technique delivers a much needed strategy for analyzing massive datasets.

Acknowledgments. We are thankful to Robert Weiss for modeling discussion and suggestions and to Eddie Holmes and Andrew Rambaut for generously providing the results from their analyses and feedback. Interested readers may request the source-code for this project from J. A. Tom.

SUPPLEMENTARY MATERIAL

Supplement A: Details of sampling from the complete model (DOI: 10.1214/10-AOAS349SUPP; .pdf). We detail the sampling steps for our complete model outlined in Section 4.3 and our constrained covariance matrices model outlined in Section 4.4.1.

REFERENCES

ALLISON, D. B., VISSCHER, P. M., ROSA, G. J. M. and AMOS, C. I. (2009). Statistical genetics and statistical genomics: Where biology, epistemology, statistics, and computation collide. *Comput. Statist. Data Anal.* **53** 1531–1534.

ANDERSON, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired* **16**. Available at http://wired.com/science/discoveries/magazine/16-07/pb_theory.

BALAKRISHNAN, S. and MADIGAN, D. (2006). A one-pass sequential Monte Carlo method for Bayesian analysis of massive datasets. *Bayesian Anal.* **1** 345–362. MR2221268

BOSCARDIN, W. J. and ZHANG, X. (2004). Modeling the covariance and correlation matrix of repeated measures. In *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives* (A. Gelman, ed.) 215–226. Wiley, Chichester. MR2138258

BUTLER, D. (2009). Swine flu goes global. *Nature* **458** 1082–1083.

CACOULLOS, T. (1964). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18** 179–189. MR0210255

CAPPÉ, O., GODSILL, S. J. and MOULINES, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *IEEE Proceedings* **95** 899–924.

CARLIN, J. B. (1992). Meta-analysis for 2×2 tables: A Bayesian approach. *Stat. Med.* **11** 141–158.

CELEUX, G., HURN, M. and ROBERT, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450

CHEN, M. H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89** 818–824. MR1294727

CHEN, S. X. (2000). Probability density function estimation using gamma kernels. *Ann. Inst. Statist. Math.* **52** 471–480. MR1794247

CHOPIN, N. (2002). A sequential particle filter method for static models. *Biometrika* **89** 539–552. MR1929161

CLERCQ, E. D. (2006). Antiviral agents active against influenza A viruses. *Nature Reviews Drug Discovery* **5** 1015–1025.

CRESSIE, N., OLSEN, A. and COOK, D. (1997). Massive data sets: Problems and possibilities, with application to environmental monitoring. In *Massive Data Sets: Proceedings of a Workshop* 115–119. National Academy Press, Washington, DC.

DELLAPORTAS, P., FORSTER, J. J. and NTZOUFRAS, I. (2002). On Bayesian model and variable selection using MCMC. *Statist. Comput.* **12** 27–36.

DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.) 3–13. Springer, New York. MR1847784

DRUMMOND, A. J., NICHOLLS, G. K., RODRIGO, A. G. and SOLOMON, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161** 1307–1320.

DRUMMOND, A. J., RAMBAUT, A., SHAPIRO, B. and PYBUS, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22** 1185–1192.

EFRON, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40** 1–11. MR0828575

EFRON, B. and MORRIS, C. (1977). Stein's paradox in statistics. *Scientific American* **236** 119–127.

FAUCI, A. S. (2005). Race against time. *Nature* **435** 423–424.

GELMAN, A. (2004). *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, FL. MR2027492

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics, vol. 4* (J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 169–193. Oxford Univ. Press, New York. MR1380276

GHEDIN, E., SENGAMALAY, N. A., SHUMWAY, M., ZABORSKY, J., FELDBLYUM, T., SUBBU, V., SPIRO, D. J., SITZ, J., KOO, H., BOLOTOV, P., DERNOVOY, D., TATUSOVA, T., BAO, Y., ST GEORGE, K., TAYLOR, J., LIPMAN, D. J., FRASER, C. M., TAUBENBERGER, J. K. and SALZBERG, S. L. (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* **427** 1162–1166.

GLYMOUR, C., MADIGAN, D., PREGIBON, D. and SMYTH, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery* **1** 11–28.

HÄRDLE, W. (1990). *Smoothing Techniques with Implementation in S*. Springer, New York.

HÄRDLE, W., MÜLLER, M., SPERLICH, S. and WERWATZ, A. (2004). *Nonparametric and Semiparameteric Models*. Springer, New York. MR2061786

HEIN, J., SCHIERUP, M. H. and WIUF, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford Univ. Press. MR2120677

HOLMES, E. C., GHEDIN, E., MILLER, N., TAYLOR, J., BAO, Y., ST GEORGE, K., GRENFELL, B. T., SALZBERG, S. L., FRASER, C. M., LIPMAN, D. J. and TAUBENBERGER, J. K. (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biology* **3** 1579–1589.

HUDSON, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7** 1–44.

KASS, R. E. and WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91** 1343–1370.

KETTENRING, J. R. (2009). Massive datasets. *Comput. Statist.* **1** 25–32.

KNOWLES, L. L. (2004). The burgeoning field of statistical phylogeography. *J. Evol. Biol.* **17** 1–10.

KUO, L. and MALLICK, B. (1998). Variable selection for regression models. *Sankhya: The Indian Journal of Statistics B* **60** 65–81. MR1717076

LAGAKOS, S. W. (2006). The challenge of subgroup analysis—reporting without distorting. *New England Journal of Medicine* **354** 1667–1669.

LAMBERT, D. (2003). What use is statistics for massive data? In *Lecture Notes—Monograph Series, Crossing Boundaries: Statistical Essays in Honor of Jack Hall* (J. E. Kolassa and D. Oakes, eds.) 217–228. IMS, Beachwood, OH. MR2125060

LIANG, L. J. and WEISS, R. E. (2007). A hierarchical semiparametric regression model for combining HIV-1 phylogenetic analysis using iterative reweighting algorithm. *Biometrics* **63** 733–741. MR2395710

LIANG, L. J., WEISS, R. E., REDELINGS, B. and SUCHARD, M. A. (2009). Improving phylogenetic analyses by incorporating additional information from genetic sequence databases. *Bioinformatics* **25** 2530–2536.

MININ, V. M., BLOOMQUIST, E. W. and SUCHARD, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25** 1459–1471.

NELSON, M. I. and HOLMES, E. C. (2007). The evolution of the epidemic influenza. *Nature Reviews Genetics* **8** 196–204.

RAMBAUT, A., PYBUS, O. G., NELSON, M. I., VIBOUD, C., TAUBENBERGER, J. K. and HOLMES, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453** 615–620.

RIDGEWAY, G. and MADIGAN, D. (2003). A sequential Monte Carlo method for Bayesian analysis of massive datasets. *Data Mining and Knowledge Discovery* **7** 301–319. MR1977337

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

RUBIN, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 395–402. Oxford Univ. Press.

SALOMON, R. and WEBSTER, R. G. (2009). The influenza virus enigma. *Cell* **136** 402–410.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York. MR1191168

SMITH, G. J. D., VIJAYKRISHNA, D., BAHL, J., LYCETT, S. J., WOROBEY, M., PYBUS, O. G., MA, S. K., CHEUNG, C. L., RAGHWANI, J., BHATT, S., PEIRIS, J. S. M., GUAN, Y. and RAMBAUT, A. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459** 1122–1125.

SUCHARD, M. A. and RAMBAUT, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25** 1370–1376.

SUCHARD, M. A., KITCHEN, C. M. R., SINSHEIMER, J. S. and WEISS, R. E. (2003). Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology* **52** 649–664.

SUCHARD, M. A., WANG, Q., CHAN, C., FRELINGER, J., CRON, A. and WEST, M. (2010). Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *J. Comput. Graph. Statist.* **19** 419–438.

TOM, J. A., SINSHEIMER, J. S. and SUCHARD, M. A. (2010). Supplement to "Reuse, recycle, reweigh: Combating influenza through efficient sequential Bayesian computation for massive data." DOI: 10.1214/10-AOAS349SUPP.

WARN, D. E., THOMPSON, S. G. and SPIEGELHALTER, D. J. (2002). Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Stat. Med.* **21** 1601–1623.

WEBSTER, R. G. and WALKER, E. J. (2003). The world is teetering on the edge of a pandemic that could kill a large fraction of the human population. *Amer. Sci.* **91** 122.

WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics* **16** 97–159.

YEWDELL, J. and GARCIA-SASTRE, A. (2002). Influenza virus still surprises. *Current Opinion in Microbiology* **5** 414–418.

ZWICKL, D. J. and HOLDER, M. T. (2004). Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Systematic Biology* **53** 877–888.

J. A. TOM
DEPARTMENT OF BIOSTATISTICS
UCLA SCHOOL OF PUBLIC HEALTH
LOS ANGELES, CALIFORNIA 90095
USA
E-MAIL: jentom@ucla.edu

J. S. SINSHEIMER
M. A. SUCHARD
DEPARTMENTS OF BIOMATHEMATICS
AND HUMAN GENETICS
DAVID GEFFEN SCHOOL OF MEDICINE AT UCLA
DEPARTMENT OF BIOSTATISTICS
UCLA SCHOOL OF PUBLIC HEALTH
LOS ANGELES, CALIFORNIA 90095
USA
E-MAIL: janet@mednet.ucla.edu
msuchard@ucla.edu