Does History Repeat Itself? Wavelets and the Phylodynamics of Influenza A

Jennifer A. Tom,*1 Janet S. Sinsheimer,2,3 and Marc A. Suchard2,3

1Department of Biostatistics, School of Public Health, University of California, Los Angeles
2Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles
3Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles

*Corresponding author: E-mail: jen.tom@gmail.com.
Associate editor: Alexei Drummond

Abstract

Unprecedented global surveillance of viruses will result in massive sequence data sets that require new statistical methods. These data sets press the limits of Bayesian phylogenetics as the high-dimensional parameters that comprise a phylogenetic tree increase the already sizable computational burden of these techniques. This burden often results in partitioning the data set, for example, by gene, and inferring the evolutionary dynamics of each partition independently, a compromise that results in stratified analyses that depend only on data within a given partition. However, parameter estimates inferred from these stratified models are likely strongly correlated, considering they rely on data from a single data set. To overcome this shortfall, we exploit the existing Monte Carlo realizations from stratified Bayesian analyses to efficiently estimate a nonparametric hierarchical wavelet-based model and learn about the time-varying parameters of effective population size that reflect levels of genetic diversity across all partitions simultaneously. Our methods are applied to complete genome influenza A sequences that span 13 years. We find that broad peaks and trends, as opposed to seasonal spikes, in the effective population size history distinguish individual segments from the complete genome. We also address hypotheses regarding intersegment dynamics within a formal statistical framework that accounts for correlation between segment-specific parameters.

Key words: phylogenetics, Bayesian nonparametrics, wavelets, importance sampling, influenza A.

Introduction

Seasonality and patterns of gene segment interaction characterize the evolutionary history of influenza A and hint at intriguing underlying patterns. In particular, we are interested in the interaction between population level factors and influenza A viral evolution (Grenfell et al. 2004). A clear understanding of this virus is vital to preventing the yearly seasonal epidemics that currently cause approximately 500,000 deaths worldwide (Stöhr 2002) and occasional pandemics that can afflict 25–30% of the global population (Taubenberger et al. 2001). Characterizing this fundamental cycle of deadly reemergence leads us to our current efforts in modeling patterns of past influenza A demographics.

Influenza A is composed of eight segments that encode 11 proteins (Nelson and Holmes 2007) and employs several evolutionary mechanisms, including antigenic drift and reassortment. Antigenic drift refers to the accumulation of point mutations on the two surface glycoproteins of haemagglutinin (HA) and neuraminidase (NA) due to a high mutation rate of approximately one every replication cycle (Nelson and Holmes 2007). There are 9 NA (N1–N9) and 16 HA (H1–H16) variants in the avian reservoir that display markedly different antigenicity (de Jong et al. 2000) with the subtype determined by the combination of the two glycoproteins. Reassortment occurs when two viral subtypes coinfect a single host cell and exchange entire segments of influenza A. This can occur between species (e.g., pig and human), subtypes (e.g., H3N2 and H1N1), or variants within a subtype (Nelson et al. 2008).

We address the compelling need for biologically motivated statistical methods that characterize the demographics arising from an interaction of segments and subtype. We focus in this present inquiry on three aspects of influenza A demographics. First, we locate the timing of genetic diversity peaks; these peaks generally occur early in the season (Creanza et al. 2010) but vary in timing (Russell et al. 2008). Second, we explore the more subtle feature of subtype interaction. There are two major seasonal influenza A subtypes, H3N2 and H1N1, that cocirculate yet exchange dominance (Holmes 2010). One explanation is that H3N2 evolution is marked by stasis during which H1N1 dominates followed by evolutionary bursts. During stasis periods, the HA segment has increased levels of genetic diversity as no single variant is dominant (Wolf et al. 2006). Finally, we determine if there are underlying frequencies in which individual segments of influenza A differ from the complete genome process.

The ancestral relationships between sampled isolates of influenza A are inferred using extant sequences that retain the imprint of past demographic events. To describe past demographic history, we rely on estimates of effective population size that reflect the genetic diversity in a population (Wright 1931) rather than the (generally larger) census count (Charlesworth 2009). Selection can cause
effective population size to vary across a single genome (Charlesworth 2009). Effective population size is more concretely described as the census population size necessary to capture the same allele frequencies under the ideal Wright–Fisher model of reproduction (Kliman et al. 2008).

The well-characterized coalescent process (Kingman 1982) parameterizes effective population size through time. The coalescent process defines the mechanism for forming the topology of the phylogenetic tree. Extant and inferred ancestral lineages coalesce at internal nodes and the distribution of the intercoalescent intervals or waiting times between nodes informs our understanding of the demographic history (Nee et al. 1995). The models for effective population size range from the parsimonious constant population size (Kuhner et al. 1995), exponential growth (Kuhner et al. 1998), and logistic growth (Pybus et al. 2001) to highly parameterized approaches whose forms depict more heterogeneous histories.

Highly parameterized coalescent-based models began with the lineages through time (LTT) plot (Nee et al. 1995) depicting intercoalescent times versus cumulative number of coalescent events and the closely related classic skyline plot (Pybus et al. 2000), which transforms the outcome of the LTT plot into effective population size. Both methods estimate parameters for all intercoalescent intervals. Strimmer and Pybus (2001) adjoin intercoalescent intervals into a predetermined number of groups conditioned on a given genealogy that the Bayesian skyline plot model (BSP; Drummond et al. 2005) extends by integrating over all genealogies and allowing for data sampled at different times (heterochronous). Flexible specification of the number of predetermined groups followed (Opplen-Rhein et al. 2005) using reversible jump Markov chain Monte Carlo (MCMC), a flexible Gaussian Markov random field prior (Minin et al. 2008), and Bayesian stochastic variable selection (Heled and Drummond 2008).

Approaches for formalized testing within the LTT and BSP framework remain unclear suggesting that their use may lie in an exploratory phase to inform the selection of a more parsimonious model (Emerson et al. 2001). For the highly irregular yet cyclical history of influenza A, the choice among previously developed models for effective population size is ambiguous leading us to explore the wavelet basis. This basis is a tool gaining traction in the statistical community (Silverman 1999; Vidakovic 1999; Abramovich et al. 2000) that affords both location specificity and periodicity and can capture rapid changes in signal such as cusps and spikes (Yang 1995; Ruggeri and Vidakovic 2005) on a seasonal scale.

Phylogenetic inference for a massive sequence data set, requiring a model space on the order of gigabytes to terabytes (Cressie et al. 1996), is becoming imaginable due to increased sequencing resources (Bao et al. 2008; Kuiken et al. 2010). In practice, however, we are limited by the computational burden that renders these complex models intractable. One suboptimal solution is stratification (Cressie et al. 1996; Kettenring 2009) or partitioning the data set into smaller stratified data sets and analyzing each independently as is now common in phylogenomic inference (Shapiro et al. 2004; Atkinson et al. 2008; Rambaut et al. 2008). In a Bayesian framework, posterior samples from the inference of these forced independent data sets constitute stratified or intermediate realizations. This strategy, while tractable, overparameterizes. This partitioned data set is best represented in a hierarchical framework (Laird and Ware 1982) that accounts for correlation among the parameters in the data partitions (Suchard et al. 2003) and comes with the additional benefit of shrinkage estimators (Efron and Morris 1977).

We estimate this hierarchical model in a two-stage process in which the first is sampling of realizations of parameters conditional only on data within that partition. In the second stage, we capitalize on the intermediate realizations available from the, already computationally intensive, stratified analyses. We previously explored this technique using a low-dimensional multivariate summary statistic of the phylogenies (Tom et al. 2010). This work extends the principal groundwork of the dynamic iteratively reweighting MCMC algorithm (DyRIMA) (Liang and Weiss 2007; Liang et al. 2009). DyRIMA relies on importance sampling (Rubin 1988; Smith and Gelfand 1992) and meta-analysis techniques that create a single hierarchical model from summary statistics inferred from independent studies. We adapt this technique in the present inquiry to an even wider range of biological questions by applying it to the branch lengths and lineages that comprise a phylogenetic tree. We infer a non-parametric hierarchical wavelet-based model with a massive data set to address epidemic timing, H3N2 and H1N1 subtype interaction, and intersegment dynamics. We begin in Wavelets in Statistics with essentials of the wavelet basis, provide the introductory section Phylogenetic Inference, describe our extension of DyRIMA in Computational Recyling, continue with a section on Model Specification, and conclude in Influenza Resources and Computing Issues with a brief review of our available stratified realizations. We then present a Results section and conclude with a Discussion section.

Materials and Methods

Wavelets in Statistics

We provide a brief outline of wavelets which is not intended as a rigorous mathematical treatment of the topic (Daubechies 1992; Strang and Nguyen 1996) or a detailed survey of applications from a statistical perspective (Silverman 1999; Vidakovic 1999; Abramovich et al. 2000). We want to represent the continuous function of log-effective population size, \( \lambda(t) \), with an orthonormal wavelet basis. The reader is likely aware of the more familiar Fourier basis that is derived from the orthogonal basis functions of sines and cosines. Similarly, the wavelet basis relies on two functions: the scaling function, \( \phi \) and the wavelet function, \( \psi \). We select the Daubechies wavelet basis, specifically the D4 filter illustrated in figure 1a in the present inquiry.

The wavelet basis is unique in that it allows multiresolution or analysis at multiple levels of detail. This key concept
enables us to capture the big picture over 13 seasons and also focus on seasonal events. Our orthogonal basis is comprised of shifts and dilations of \( \phi \) and \( \psi \) via the equations

\[
\phi_k(t) = \phi(t - k) \quad \text{and} \quad \psi_{jk}(t) = 2^j \psi(2^j t - k),
\]

where \( j \) and \( k \) are integers. The index \( j \) controls the scale or resolution and \( k \) controls location. We represent our function as

\[
\lambda(t) = \sum_k c_k \phi_k(t) + \sum_{j, k} d_{jk} \psi_{jk}(t),
\]

where \( c_k \) and \( d_{jk} \) are the scaling and wavelet coefficients. We solve for these coefficients using the inner products

\[
c_k = \int \lambda(t) \phi_k(t) \, dt \quad \text{and} \quad d_{jk} = \int \lambda(t) \psi_{jk}(t) \, dt.
\]

In practice, we only need to evaluate or observe \( \lambda(t) \) at a finite number of equally spaced timepoints \( \omega_z \), where \( z = (1, \ldots, G \equiv 2^l) \) is an integer and

\[
\log \Theta = (\log \omega_1, \ldots, \log \omega_G) = (\log \theta_1, \ldots, \log \theta_C).
\]

Our total number of timepoints is \( G \) and we refer to our observed vector of data points as \( \log \Theta \). We now want to transform \( \log \Theta \) into wavelet space. This is achieved in practice using the discrete wavelet transform (DWT), described in Strang and Nguyen (1996), that relies on the orthogonal matrix \( W \) of dimension \( G \times G \). The entries in \( W \) rely on the chosen wavelet basis and Strang (1992) provides a clear derivation of the Daubechies D4 coefficients. The DWT of our observed log-effective population sizes, \( W \log \Theta = \beta \), yields the vector of scaling and wavelet coefficients, \( \beta \). Due to orthogonality, the inverse discrete wavelet transform (IDWT) is simply the transpose of the DWT so that

\[
\log \Theta = W' \beta.
\]

The DWT has a fast implementation of \( O(G) \) due to Mallat’s pyramid algorithm (Mallat 1989).

The vector of coefficients, \( \beta = (c_0, d_{11}, d_{21}, d_{31}, d_{32}, d_{33}, d_{34}, \ldots) \), warrants closer inspection as it reflects our ability to analyze at multiple levels of resolution. Note first we have \( j \) total resolution levels. Within a given level \( j \), there are \( K_j = 2^{j-1} \) wavelet coefficients \( d_{jk} \). This means our total number of coefficients is \( 1 + \sum_{j=1}^{\log_2 G} 2^{j-1} \) that includes the
single scaling function \( c_0 \) added onto the wavelet coefficients. We will refer to both the scaling and wavelet coefficients as wavelet coefficients from now on. Each increasing level of detail from \( j \) to \( j + 1 \) is captured by twice as many coefficients. At level \( j = 1 \), we have \( d_{11} \) corresponding to a single wavelet function. At \( j = 2 \), we compress our function so both \( d_{21} \) and \( d_{22} \) cover our signal length and so on.

We clarify the effects of these coefficients now with the motivational example depicted in figure 1b. This signal represents 16 years of log-effective population size data for a single subtype combined for the Northern and Southern hemispheres. We see two strong signals per season and a spike every four seasons (at 2, 10, and 14) followed by a rapid decrease in signal. We also have an anomalous period with a constant log-effective population size corresponding to a season with another dominant subtype. In figure 1c, we show a reconstruction of the signal based on different numbers of wavelet coefficients. We show a reconstruction (solid line) using the first three wavelet levels for eight total coefficients. This allows us to capture the large spikes at seasons 2, 10, and 14. However, we need to also include the next three wavelet levels, corresponding to 64 total coefficients or 32 at the highest level of detail, in order to capture the half-seasonal effects (dashed line).

We visually assess our wavelet coefficients with scalograms (Alsberg et al. 1998). These depict the absolute magnitude of the wavelet coefficients rescaled to lie between 0 (white, small) and 1 (black, large). Each level of detail from \( j \) to \( j + 1 \) is captured by twice as many coefficients as wavelet coefficients from now on. Each increasing level of detail from \( j \) to \( j + 1 \) is captured by twice as many coefficients. At level \( j = 1 \), we have \( d_{11} \) corresponding to a single wavelet function. At \( j = 2 \), we compress our function so both \( d_{21} \) and \( d_{22} \) cover our signal length and so on.

**Phylogenetic Inference**

We next review the reconstruction of a phylogenetic tree that is described elsewhere in more detail from both a frequentist (Felsenstein 1981) and a Bayesian perspective (Rannala and Yang 1996; Sinsheimer et al. 1996, 1997; Mau et al. 1999; Li et al. 2000). A phylogeny describes the evolutionary relatedness of sampled sequences using a genealogy, \( g \), comprised of nodes and branches. We focus on serially sampled taxa under a coalescent process (Rodrigo and Felsenstein 1999) that allows the coalescent time on the tree to be rescaled into calendar time.

In the heterochronous case, there are two possible events of interest, sampling and coalescence, for a total of \( E = N + S - 2 \) events, where \( N \) is the total number of taxa and \( S \) is the number of sampling times. These events are indexed by \( e = (1, \ldots, E) \). We denote present day as \( t_0 = 0 \) and form the genealogy by randomly selecting two of the taxa at time \( t_{e-1} \) to coalesce further back in time at \( t_e \). The intercoalescent time \( u_e = t_e - t_{e-1} \) or time between coalescent events \( u_1, \ldots, u_E \) are independent exponential random variables. We assume that there is an effective population size vector \( \Theta = (\theta_1, \ldots, \theta_E) \) that is a discrete representation of the effective population size through time and spans the entire inferred time of the phylogeny.

The effective population sizes can change at \( G - 1 \) times where \( G < E \) and \( G \) is specified in advance. Rambaut et al. (2008) employ the BSP model (Drummond et al. 2005) that groups adjoining intercoalescent intervals, so estimates of effective population size are based on multiple intervals. The locations of the \( G - 1 \) change points are inferred in a Bayesian framework resulting in smoothing of the effective population size plots. In contrast, the wavelet basis requires equally spaced time points or a regularized grid placed on the effective population sizes. This means effective population size is no longer constrained to change only at coalescent events. The number of bins \( G \) is set so the location of our change points or grid times are \( w = (w_1, \ldots, w_G) \). The time elapsed between grid times is \( \Delta w = w_2 - w_1 \).

The implication of this grid for the likelihood of the genealogy is that the effective population size can change between events \( e \). Let \( b = 1, \ldots, B \) index the combined grid and event times that are unique. Let \( \delta_b \) be the elapsed times between the grid and event times and \( k_b \) the number of lineages during the corresponding time interval. Then the serial coalescent likelihood of the genealogy is

\[
P(g|\Theta) = \prod_{b=1}^{B} \left( \frac{\kappa_b (k_b - 1)}{2\theta_{h(b)}} \right)^{1_{\text{coa}(b)}} \exp \left( \frac{\kappa_b (k_b - 1) \delta_b}{2\theta_{h(b)}} \right),
\]

where \( 1_{\text{coa}(b)} = 1 \) indicates a coalescent event occurred at event or grid time \( b \) and \( 1_{\text{coa}(b)} = 0 \) indicates it did not (Rodrigo and Felsenstein 1999; Drummond et al. 2002). The mapping \( h(b) \) which maps \( b \) to \( z \) is \( h(b) = z \) if \( z > b \delta_a \leq z \Delta w \) and \( \sum_{a=1}^{b-1} \delta_a > (z - 1) \Delta w \) (refer to fig. 2).

**Computational Recycling**

Rambaut et al. (2008) partitioned their influenza sequence data into strata and independently inferred a phylogeny for each strata. Let the symbol \( \Omega \) represent all the parameters in the given model, \( \Omega_{\text{BSP}} \) the parameter specification under the BSP model and \( \Omega_{\text{WAV}} \) the parameters under the nonparametric Bayesian hierarchical wavelet model. Rambaut et al. (2008) generated intermediate realizations during estimation of these stratified phylogenies or in other words samples from \( P(g|D, \Omega_{\text{BSP}}) \), which we will weigh. DyLRMA (Liang and Weiss 2007) facilitates this process and ultimately substitutes \( P(g|\Omega_{\text{WAV}}) \) for the prior \( P(g|\Omega_{\text{BSP}}) \) defined in equation (2). This weighing allows us to estimate the joint hierarchical model \( P(g|D) \) using the joint prior \( P(g|\Omega_{\text{WAV}}) \).

Let us assume that given a genealogy the sequence data are independent of the model parameters or \( P(D|g, \Omega) = P(D|g) \). Let \( i = (1, \ldots, l) \) be the index
where $P$ specification and de-emphasize those that are very likely hierarchical or weigh stratified realizations with a high likelihood under the grid the mapping $h$ events that occur at times $t_i$. There are $N + 5 - 2 = 6$ or $3 + 2 = 5$ events that occur at times $t_1, \ldots, t_9$. Effective population size changes at regularly spaced intervals so can change at $w_i, w_j$, and $w_k$ with the depicted grid $G = 4$. On the right-hand side, there are nine unique event and grid times indexed by $b$ along with the corresponding elapsed times $\delta_b$ and the mapping $h(b)$ back to the effective population size index $z$.

For the exchangeable segment or stratified analyses. Then

$$P(g|D) = \int P(g, \Omega_{WAV}|D) \, d\Omega_{WAV}$$

$$\propto \int P(D|g, \Omega_{WAV})P(g|\Omega_{WAV}) \times P(\Omega_{WAV}) \, d\Omega_{WAV}$$

Since $P(g|D_i, \Omega_{BSP}) \propto P(D_i|g)P(g|\Omega_{BSP})$, we have

$$P(g|D) \propto \int \prod_{i=1}^{9} P(g|D_i, \Omega_{BSP}) \frac{P(g|\Omega_{WAV})}{P(g|\Omega_{BSP})} \times P(\Omega_{WAV}) \, d\Omega_{WAV}.$$  

where $P(g|D_i, \Omega_{BSP})$ are the realizations generated under the $i$th stratified analysis. We identify our importance weights as:

$$w(g, \Omega_{WAV}) = \frac{P(g|\Omega_{WAV})}{P(g|\Omega_{BSP})},$$

which intuitively make sense as we emphasize or highly weight stratified realizations with a high likelihood under the hierarchical or $\Omega_{WAV}$ specification relative to the $\Omega_{BSP}$ specification and de-emphasize those that are very likely under the stratified or $\Omega_{BSP}$ specification relative to the $\Omega_{WAV}$ specification.

Of practical importance is the ability to sample from the conditional distribution of a genealogy for a given stratified analysis under the joint model, $P(g|D, \Omega_{WAV})$, which is based on importance sampling machinery (Rubin 1988) and shown in Liang and Weiss (2007) as:

$$P(g|D, \Omega_{WAV}) \Delta_{IRMA}$$

$$= \frac{1}{W_i} \sum_{m=1}^{M} w(g^{(m)}, \Omega_{WAV}) \delta_{g^{(m)}}(g).$$

The function $W_i = \sum_{m=1}^{M} w(g^{(m)}, \Omega_{WAV})$ normalizes the weights and $\delta_{g^{(m)}}(g)$ is a degenerate distribution at $g^{(m)}$. On a practical note, we sample from our posterior distributions as follows: At the $m$th iteration, we sample $g^{(m)}$ using equation (6). The numerator of the weights (identified in eq. 5) needs to be updated during each round of MCMC and is $P(g^{(m)}|\Omega_{WAV})$. The denominator of the weights, $P(g^{(m)}|\Omega_{BSP})$, is calculated using the prior distribution specified in the stratified model. Now, having sampled $g^{(m)}$, we condition on this value to update the parameters $\Omega_{WAV}$ and repeat this process for a total of $M$ iterations. We now have all the statistical machinery our model requires for estimation and proceed to model specification.

**Model Specification**

We explore a nonparametric hierarchical framework to capture the peaks and valleys in effective population size
corresponding to genetic diversification and selective sweeps that punctuate the historic landscape of influenza A. The demographic history of influenza A is periodic as there is an influenza A season every winter. We are interested in addressing our biological questions of interest while accounting for sudden peaks and valleys that punctuate the periodic evolutionary history of influenza A. To explore the rich and varied history of influenza A, we use a wavelet basis $W$. Let $\Theta_i$, where $i = (1, \ldots, l = 8)$, be the effective population size for a given influenza A segment. We model using a log link as the effective population sizes are strictly positive. Our exploratory wavelet model is

$$\log \Theta_i = \mu_i + W^\prime \eta_i. \quad (7)$$

Each segment has its own mean log-effective population size $\mu_i$ and $\eta_i$ is the vector of wavelet coefficients, where $c_{00}$, or the scaling coefficient that captures the overall mean, is not inferred to avoid overparameterization. Priors are specified as $\mu_i \sim N(\mu_0, \sigma_\mu^2)$ where $\mu_0 \sim N(m_{i0}, \sigma_{m0}^2)$ and $\sigma_\mu^2 \sim \text{Inv-Gamma}(a_{\mu}, b_{\mu})$. The $l \times 1$ vector $\mu = (\mu_1, \ldots, \mu_l)$ contains all the log mean segment effects and the hyperprior constants $m_{i0}, \sigma_{m0}$, and $b_{\mu}$ are noninformative. Each segment has its own wavelet process captured with $W^\prime \eta_i$. The $K^j \times 1$ vector $\eta_i$ is distributed $\eta_i \sim N(\eta_0, \Sigma_K (\nu^2_j, \lambda_j))$. Coefficients of a given level $j$ are assumed to have the same mean; the multivariate normal $\eta_i$ has a mean $\eta_0, \Sigma_K$, where $\Sigma_K$ is a $K^j \times 1$ vector of ones. The function $\Sigma(\sigma^2, \rho)$ forms an auto-regressive (AR1) covariance matrix. In an AR1 matrix, given the covariance matrix is indexed by $k_1 = (1, \ldots, K)$ and $k_2 = (1, \ldots, K)$, element $(k_1, k_2) = \sigma^2 \rho^{k_1 - k_2}$. Therefore, $\Sigma(\nu^2_j, \lambda_j)$ is an AR1 matrix of size $K^j \times K^j$ where each segment $i$ and level $j$ has its own variance $\nu^2_j$ and correlation $\lambda_j$. The wavelet coefficients are independent between levels $j$ and correlated within a level with the strength of correlation decreasing with increased distance between wavelet coefficients as in Vidakovic and Müller (1995). To complete specification, $\eta_0, \Sigma_K, \Sigma(\nu^2_j, \lambda_j)$, $\nu^2_j \sim \text{Inverse-Gamma}(a_v, b_v)$, $\lambda_j \sim \text{Beta}(a_\lambda, b_\lambda)$, and the hyperprior constants $m_{i0}, s_{i0}, \sigma_{i0}, b_{i0}, a_{i0}$, and $b_{\mu}$ are noninformative.

Next we examine segment-specific frequencies by determining, for example, frequencies in which glycoprotein NA and the nonstructural protein NS differ from the complete genome mean process. Our strategy relies on Gibbs variable selection (Kuo and Mallick 1998) applied to entire segment-specific wavelet levels. We refer to this model, as our joint wavelet model. The mean landscape of the segment log demographic history is described by the wavelet process $W^\prime \beta$, where $\beta_2 \sim N(\beta_0, 1K^j, \Sigma (\sigma^2_j, \rho_j))$. The AR1 covariance matrix, $\Sigma (\sigma^2_j, \rho_j)$, is $K^j \times K^j$ with $\sigma^2_j$ as the variance and $\rho_j$ as the correlation. Furthermore, $\beta_0 \sim N(\beta_{10}, s_{10})$, $\sigma^2_j \sim \text{Inverse Gamma}(a_{\sigma}, b_{\sigma})$, and $\rho_j \sim \text{Beta}(a_{\rho}, b_{\rho})$, where $m_{10}, s_{10}, \sigma_0, b_{\sigma}, a_{\sigma},$ and $b_{\sigma}$ are noninformative. Segment-specific frequencies are selected by level as parameters in the matrix $\Gamma_1$ perform variable selection on entire frequency levels. The introduction of Gibbs variable selection on wavelet coefficients was previously outlined in Clyde et al. (1998). This is a $G \times G$ diagonal matrix such that $\Gamma_1 = \text{diag}(\gamma_{mn})$, $x = j$, where $x = (2, \ldots, G)$ if $x > 2^{j-1}$ and $x \leq 2^j$, $n_1 = 0$, and $\gamma_{mn} \sim \text{Bernoulli}(p_0 = 0.5)$. Notice we do not estimate $\gamma_{00}$ and there must be a restriction on $\gamma_{mn}$, so that if all indicators in a given level are selected the mean process is not also inferred to avoid overparameterization.

### Influenza Resources and Computing Issues

We reexamine the stratified analyses and their resulting realizations compiled by and described in Rambaut et al. (2008). In brief, Rambaut et al. (2008) partitioned 1,302 complete genome influenza A sequences spanning 13 seasons by geographical location, subtype, and genomic segment. Each partition took approximately 2–3 weeks to infer a phylogeny on high-end computers. Parameters estimated in these phylogenetic models are used to generate smoothed BSPs that depict effective population size through time. We explore a subset of 687 H3N2 sequences New York (NY). We use all eight segments of influenza A, indexed by $i = (1, \ldots, l = 8)$, which comprise the complete genome of influenza A. There are 12 seasons available ranging from 1993 to 2005 with data missing for 2001 due to H1N1 dominance. We reduce the 18,000 available intermediate realizations to 1,000 to balance adequate coverage against efficiency.

A proper level of resolution and signal padding enables the wavelet coefficients to capture the effect of a single season (assumed as six months). We increased the total elapsed seasons for NY from 13 to 16 as it is the next lowest power of 2 that allows us to use all 13 years of data from NY. We select a grid size $G = 64$ or number of wavelet levels $j = 6$, which provides two wavelet coefficients per year at the highest level of detail. This grid size addresses our need for parsimony and a biologically interesting level of resolution. All models combine three independent chains that ran for $10^7$ iterations, discarded $10^6$ burn-in, and retained 10% of the iterations after thinning.

### Results

We examine results from the exploratory wavelet model described in equation (7). Note that we are not taking full advantage of the DyIRMA methodology as this model provides no intersegment joint parameters. In figure 3, which depicts HA for subtype H3N2 in NY, the log-effective population size fit for the exploratory wavelet model (refer to fig. 3(b)) closely follows the BSP model (refer to fig. 3(a)) with a single exception. In the 2001 season, known to be H1N1 dominant, the posterior mean of the wavelet model has a sharply defined drop in log-effective population size. In contrast, the posterior mean of the BSP model displays a slowly declining peak due to an artifact of oversmoothing. Note also the highest peak for HA in each interval occurs in the first quarter of the year (January–March) in 9 out of 12 years (excluding 2001). We next comment on the greater diversity of HA prior to the H1N1 dominant seasons 2001 and 2003 (Wolf
We observe strong support for this claim since the highest peaks occur in the 2000, 2002, and 2005 seasons (refer to fig. 3a and b). The 2000 and 2002 peaks occur prior to H1N1 dominant years and since we now know the 2006 season was H1N1 dominant (Nelson et al. 2008); the 2005 peak is consistent.

In figure 4, we look at scalograms (Alsberg et al. 1998), which make a useful exploratory analysis tool. Recall that level \( j = 6 \) has 32 coefficients with 8 serving as padding and correspond to half-seasonal effects. Similarly, level 5 are seasonal effects, level 4 are biennial effects, and so on. The first tile depicted in level 5 refers to \( \eta_{51} \) and denotes the period of time between 1 April 1993 (1993.25) and 1 April 1994 (1994.25). The years are shifted by one quarter to correspond to the most recent influenza A sampling time of 1 April 2005 (2005.25). Looking at HA in figure 4a, we see a peak persisting through all levels of detail prior to the 1995 season. In contrast, NP, in figure 4b, shows a very broad peak in that same time period but a narrow peak in 1999. We confirm the period of stasis observed in HA between 1999 and 2002 (Wolf et al. 2006) in wavelet level 4 (biennial effects) just after the 1999 season until after 2003. The evidence for this period of stasis is not as evident in NP. We do not see evidence for biennial trends that were observed (but not shown to be significant) in Finkelman et al. (2007).

One reason for this is within the H3N2 subtype, Finkelman et al. (2007) has much stronger evidence for biennial effects in the southern hemisphere and not the northern hemisphere that we are examining in the present inquiry. There are clearly no repetitive patterns; this heterogeneity emphasizes the importance of the wavelet basis.

We next explore the joint wavelet model described in equation (8) to test which frequencies in the segment-specific demographic history of influenza A differ from the complete genome. Figure 5 plots the posterior mean for \( \gamma_{in} \) across all eight segments of influenza A; these posterior means correspond to the posterior probabilities of model inclusion at each level. If the posterior probability of inclusion for a given segments wavelet levels coefficients is high, we call this level “segment specific.” Inclusion reflects that it is necessary to include that level’s wavelet coefficients in order to distinguish that segments evolutionary process them from the mean process. All segments retain the first-level wavelet coefficient \( d_{11} \) and in general do not retain yearly \( (\eta_{60}) \) and half-yearly \( (\eta_{60}) \) parameters. This suggests at increasing levels of detail, segments have similar microtrends to each other, but at broader levels of time (namely two years or more), retain segment-specific processes. One exception is NA that shows weak support for half-yearly level of details. The posterior probability of \( \gamma_{N6} \) wavelet coefficients being in the model is 0.16 or in other words, \( P(\gamma_{N6} = 1|D) = 0.16 \) and the Bayes factor (BF) of 0.32 is not extreme. NA also has the highest posterior probability of \( \gamma_{N3} \) being in the model, corresponding to 4-year blocks of time, where \( P(\gamma_{N3} = 1|D) = 0.61 \) (BF = 1.22, slight support). Looking at the scalogram in figure 4c, we see this broad peak in the last four seasons (2001–2005). We postulate that the complete genome process is driven by the unique evolution-
ary pressures of HA, which is evidenced by the low posterior probability of wavelet level indicators (listing $\gamma_{jHA}$ in order of increasing detail: 1, 0.17, 0.29, 0.07, 0, and 0.01). On the other hand, the wavelet level indicators for NA (again in order of increasing detail: 1, 0.09, 0.61, 0.09, 0.0, and 0.16) reflect the differences from the phylogenetic history of the HA segment (Nelson et al. 2006; Rambaut et al. 2008). Also apparent from figure 5 is the shrinkage of our wavelet coefficients toward zero, which is advantage of using a multiple shrinkage estimator (Clyde et al. 1998).

Finally, we again look at the posterior mean and Bayesian credible intervals (BCIs) of segment HA in figure 3c. This figure illustrates the reduced coverage of the posterior distribution of log-effective population size with the obvious exception of the 2001 season. We also see peaks every 2–5 years (in the mean process seasonal or level 5 effects, figure 4d), which aligns with the observation of Nelson and Holmes (2007) that influenza A mortality rates peak every 2–5 years in the northern hemisphere. Finally, note the complete genome process (refer to fig. 4d) exhibits clear heterogeneity that solidifies the importance of long-term trends and temporal specificity when modeling the demographics of influenza A.

**Discussion**

We extend DyRMA machinery and efficiently combine intermediate realizations from stratified Bayesian analyses into a joint hierarchical model. This approach is necessary given estimation for each stratified analyses takes 2–3 weeks of computing time, rendering estimation of the joint hierarchical model intractable by traditional computational methods. This joint model obviates the overparameterization from the stratified analyses of individual segments of influenza A induced because the parameters describing the effective population size histories are strongly correlated. In the absence of evolutionary mechanisms, such as reassortment and antigenic drift, the segments from a single-complete genome would have parallel phylogenetic histories. Our methodology accounts for the correlation present despite these mechanisms and allows us to assign statistical significance in a formal framework. We contribute statistical tools to accommodate the large scale next-generation sequencing and resequencing efforts currently underway. The computational issues described in this current study for the influenza A viruses are small when compared against the areas we are currently exploring including the LANL HIV database (Kuiken et al. 2010) and the
hemispheres as 165 days (expected is 167 based on results in data set of northern (NY) and southern (New Zealand) posterior mean of the offset between an expanded sequence based model (results not shown) and established the pos-
1000 genomes project (1000 Genomes Project Consortium 2010) for the human genome.

The studious reader will question the necessity of this undertaking as our results may only confirm those in the stratified analyses already completed by Rambaut et al. (2008). The benefits of this more principled approach are clear as we find evidence of shrinkage estimators, reduced posterior distribution coverage, and can conduct interseg-
Posterior Probability of Wavelet Level Effect Inclusion

1 2 3 4 5 6
Wavelet Level

PB 2
PB 1
PA
NS
NA
MP
HA
N P
N S
P B 2
PB 1
PA
NS
N S
N P
MP
HA

parameters, given densely serially sampled data. This sam-
pling results in data sets of massive size but also with chal-
lenging multivariate complexity. Wavelets provide a sparse
representation of complex signals as it is easy to thresh-
old levels of, and even individual (Abramovich et al. 1998),
wavelet coefficients. One must consider that we are looking
for anomalous events within data known to display periodic
behavior. The wavelet basis captures the periodicity inher-
ent in the demographics of influenza A but is also capable
of teasing out anomalies—these anomalies are what lead to
variable subtype dominance and potential pandemics. This
approach is admittedly hampered by a lack of readily avail-
able inference tools (Grinsted et al. 2004), a shortcoming we
are currently rectifying.

Our results address both decade-long patterns and sea-
sonal events. We observe epidemic peak timing occurs late
in the season for HA and strong evidence supporting the
previous hypothesis that H3N2 HA genetic diversity is high
prior to an H1N1 dominant season. By using variable se-
lection on wavelet levels, we find that segment-specific
processes reveal themselves in trends spanning multiple sea-
sons. All signals, however, evince broadly defined increased
genetic diversity starting prior to the 1998 influenza A sea-
son and ending in 2005. A possible explanation is the intro-
duction of the second-generation NA inhibitor Tamiflu in
1999. This is contrary to the belief that the introduction of a
cure will decrease genetic diversity (van Ballegooijen et al.
2009). Further exploration is warranted and will be aided by
the current prodigious global influenza A surveillance effort
(Bao et al. 2008), an undertaking that will undoubtedly re-
quire novel statistical methodology developed for massive
data sets.

Acknowledgements

We are thankful to Edward C. Holmes and Andrew Rambaut
for generously providing the results from their previous strati-
fied analyses. We also thank Donatello Telesca for in-
valuable guidance with the statistical modeling. Interested
readers may request the source code for this project from
J. A. Tom. This work was supported by National Institute
of Health grants to HG002536, GM008185, GM086887, and
GM053275.


