Biostatistics (2017), **0**, 0, *pp.* 1–40
doi:10.1093/biostatistics/[upload]gel˙main˙text

Estimating AutoAntibody Signatures to Detect Autoimmune Disease Patient Subsets

ZHENKE WU*[,1], LIVIA CASCIOLA-ROSEN[2], AMI A. SHAH[2],

ANTONY ROSEN[2], SCOTT L. ZEGER[3]

[1] *Department of Biostatistics and Michigan Institute of Data Science, University of Michigan, Ann Arbor, Michigan 48109*

[2] *Division of Rheumatology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, 21224*

[3] *Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205*

*zhenkewu@umich.edu

Summary

Autoimmune diseases are characterized by highly specific immune responses against molecules in self-tissues. Different autoimmune diseases are characterized by distinct immune responses, making autoantibodies useful for diagnosis and prediction. In many diseases, the targets of autoantibodies are incompletely defined. Although the technologies for autoantibody discovery have advanced dramatically over the past decade, each of these techniques generates hundreds of possibilities, which are onerous and expensive to validate. We set out to establish a method to greatly simplify autoantibody discovery, using a pre-filtering step to define subgroups with similar specificities based on migration of labeled, immunoprecipitated proteins on sodium dodecyl sulfate (SDS) gels and autoradiography [**G**el **E**lectrophoresis and band detection on **A**utoradiograms (GEA)]. Human recognition of patterns is not optimal when the patterns are complex or scattered across many samples. Multiple sources of errors - including irrelevant intensity differences and warping of gels - have challenged automation of pattern discovery from autoradiograms.

In this paper, we address these limitations using a Bayesian hierarchical model with shrinkage

*To whom correspondence should be addressed.

priors for pattern alignment and spatial dewarping. The Bayesian model combines information from multiple gel sets and corrects spatial warping for coherent estimation of autoantibody signatures defined by presence or absence of a grid of landmark proteins. We show the preprocessing creates better separated clusters and improves the accuracy of autoantibody subset detection via hierarchical clustering. Finally, we demonstrate the utility of the proposed methods with GEA data from scleroderma patients.

Key words: Autoantibody signatures; Batch effect; Bayesian image registration; Clustering; Gel electrophoresis; Peak detection; Markov chain Monte Carlo; Measurement error; Scleroderma.

1. Introduction

Discovering disease subgroups that share distinct disease mechanisms is fundamental to disease prevention, monitoring and treatment. For example, in autoimmune diseases, specific autoimmune responses are associated with distinct disease phenotypes and trajectories (Rosen and Casciola-Rosen, 2016). Defining the molecular markers of these subgroups has value, as these markers are of diagnostic and prognostic significance, and guide management and therapy. For example, an immune response to RNA polymerase III in scleroderma is associated with cancer; this immune response arises in response to a mutation in RNA polymerase III in that patient's cancer. While many prominent specificities recognized by the immune response have been defined, many remain to be discovered. Although modern measurement technologies are revolutionizing the ability to define specificities, each technique results in hundreds of possibilities, which are onerous and expensive to validate. A simple technique identifies patterns of antibody reactivity based on the abundance of different weighted autoantigens immunoprecipitated by patient sera. Defining similar reactivity patterns prior to applying one of the new discovery technologies would greatly simplify validation and therefore the cost and speed of antigen identification.

To obtain a patient's *unknown* component autoantibodies present in serum, scientists mix

serum collected from each patient with radiolabeled lysates made from cultured cells. These lysates contain a representation of all the proteins expressed in that cell type. Antibodies in each patients serum recognize and bind tightly to the specific protein(s) in the lysate against which they are directed (termed immunoprecipitation). After further processing, electrophoresis is used to sort the immunoprecipitated mixture of molecules using a crosslinked polymer or gel that separates the proteins by weight. Because different weighted molecules move through the gel with differential speeds, the sorted molecules form distinct autoradiographed bands along the gel. By design, one gel can sort multiple samples on parallel lanes. Such experiments, referred to as gel electrophoresis autoradiography (GEA), serve to identify subsets of samples that share one or more interesting observed bands. It is noteworthy that the lysate proteins are present in their native conformation. In our experience, many autoantibodies have epitopes that are conformationally dependent, making this a powerful advantage of this method over many of the new peptide-based (linear epitopes) sequencing technologies.

In this paper, we focus on estimating a multivariate binary autoantibody signature for each sample, that represents the presence or absence of autoantibodies by their weights. We discretize the molecular weight scale (kDa) into landmarks for the signature estimation purpose and fast Bayesian image dewarping (Section 2.3.2).

To infer patient subsets, we cluster patients based upon the presence or absence of each band as well as other curve features such as the peak scale and amplitude. There are two critical barriers to the successful implementation of this approach that we address. First, there are *batch*, or *gel effects* in the raw GEA data. Ideally, identical weighted molecules should travel the same distance through the gels. This distance however varies by gel due to differential experimental conditions. Second, gels are frequently slightly warped as they electrophorese due to heating effects generated during the electrophoresis procedure and due to artifacts introduced during physical processing of the gels. As the size and complexity of GEA experiment database grows, the need for systematic,

reproducible, scalable error correction has also grown.

In this paper, we introduce and illustrate a novel statistical approach based on hierarchical Bayesian modeling with shrinkage priors for preprocessing the GEA data and estimating autoimmune disease subgroups. We focus on clustering individuals into a small number of subgroups within which people share similar autoantibody profiles estimated from the data.

We first preprocess the data by peak detection and batch-effect correction that set the stage for cross-sample comparisons. In particular, we identify the locations along the gel where the radioactive intensity rise above its neighboring level. We propose a computationally-efficient local scoring algorithm that performs well for minor peaks (Section 2.2). Guided by the detected peaks, we align and dewarp the images. Specifically, we first align multiple GEA images using piecewise linear stretching/compression anchored at marker bands on the reference lanes loaded on all the gels (Section 2.3.1). We then propose and fit a hierarchical Bayesian model that characterize spatial gel deformations approximated by tensor-product B-spline bases. We use Markov chain Monte Carlo to estimate the warping functions, the reverse application of which then dewarps the gel images. In our framework, the dewarping accuracy depends on the resolution of the discretized molecular weight landmarks (Section 2.3.2) and the pattern of detected peaks (Step 2, Appendix S4). The Bayesian framework has the advantage of incorporating inherent uncertainty in assigning a peak to a molecular landmark.

Finally, based on the aligned intensity profiles, one vector per sample lane, we use hierarchical clustering to create nested subgroups. For N samples, hierarchical clustering produces a dendrogram that represents a nested set of clusters. Depending on where the dendrogram is cut, between 1 and N clusters result. We then demonstrate through real data that preprocessing better separates clusters and improves the accuracy of cluster detection compared to naive analyses done without alignment.

At each iteration of the MCMC sampling, we can dewarp the gels and obtain the multivariate

signatures. We obtain a collection of dendrograms by hierarchically clustering the aligned intensity profiles at each iteration. In this paper, we focus on *maximum a posterior (MAP)* peak-to-landmark matching and use multiscale bootstrap resampling (Shimodaira *and others*, 2004) to assess the structural uncertainty of the dendrograms. A future work will focus on representations of statistical uncertainties using a large number of posterior dendrograms, for example, building on the log maps from metric tree spaces to Euclidean space (e.g., Billera *and others*, 2001; Willis and Bell, 2016).

The rest of the paper is organized as follows. Section 2 introduces the importance of pre-processing GEA data followed by algorithmic details for peak detection in Section 2.2 and batch effect correction in Section 2.3. In Section 3, we describe model posterior inference by MCMC and the statistical property of shrinkage priors. We demonstrate how the proposed methods function through an application to signature estimation and subgroup identification of scleroderma patients in Section 4. The paper concludes with a discussion on model advantages and opportunities for extensions.

2. Data Pre-Processing

2.1 *GEA Data and Preprocessing Overview*

Gel electophoresis for autoantibodies (GEA) is designed to separate autoantigen mixtures according to molecular weight and to radioactively map them as bands along the gel. Figure 1(a) shows one example of raw GEA image. We tested four sets of samples from scleroderma patients with a malignancy; of note, these sera were pre-selected as being negative for the 3 most commonly found scleroderma autoantibodies (anti-topoisomerase 1, anti-centromere and anti-RNA polymerase III antibodies, which in aggregate are found in $\sim 60\%$ of scleroderma patients). Each sample set consisted of 19 patient sera plus one reference comprising a mixture of protein standards with known molecular weights, referred to as *marker molecules*. The middle panel of Figure

1(b) shows marker molecule reference (lane 1) and a set of 19 patient samples (lanes 2-20), each showing the band patterns that read out autoantibodies present in that patient's serum. The intensity curves are overlaid above the heatmap in Figure 1(b). Seven clear spikes indicated by vertical lines mark the locations of the marker molecules, or *marker peaks*, from reference lane 1. The marker molecular weights decrease from the left to the right (bottom of Figure 1(b)). Using polynomial interpolation, we infer the intermediate molecular weight of a peak appearing at an arbitrary location.

Identical marker molecules scatter horizontally (empty circles, bottom panel of Figure 1(b)) caused by different experimental conditions such as the strength of the electric field. We correct the marker peak misalignment by aligning the marker peaks across gels and piecewise-linearly stretch or compress each gel anchoring at the matched marker peaks, a technique first used in human motion alignment anchored at body joints (e.g., Uchida and Sakoe, 2001).

The autoradiographic process is also vulnerable to smooth non-rigid, spatial gel deformation. This is most evident from the bands of actin, a ubiquitous protein of molecular weight 42 kDa, present in all lanes at around 0.45 (middle panel of Figure 1(b)). The bands form a smooth curve top-to-bottom. The curvature represents the gel deformation since actin has identical weight and should appear at identical locations across the 19 lanes. Without correction, this deformation interferes with accurate cross-sample assessment of the presence or absence of the *same* autoantibody. In Section 2.3.2, we propose a Bayesian hierarchical image-dewarping model to correct the deformation and align the peaks.

Let $(\mathbf{t}^0, \mathbf{M}^0) = \left\{ \left(t_b^0, M_{gib}^0 \right) \right\}$ represent the standardized, high-frequency GEA data, for bin $b = 1, \ldots, B$ on lane $i = 1, \ldots, N_g$ from gel $g = 1, \ldots, G$. Appendix S1 describes standardization of raw data. Here $\mathbf{t}^0$ is a grid that evenly splits the unit interval $[0, 1]$ with $t_{gb}^0 = b/B \in [0, 1]$. M_{gib}^0 is the radioactive intensity scanned at t_b^0 for lane $i = 1, \ldots, N_g$ on gel $g = 1, \ldots, G$. Let $N = \sum_g N_g$ be the total sample size.

For the rest of this section, we take the high-frequency data $(\mathbf{t}^0, \mathbf{M}^0)$ and map it to multivariate binary data $\mathbf{D}$ on a coarser common grid across gels. In Section 2.2, we propose a generic method to transform an arbitrary high frequency, nearly continuous intensity data into raw peak locations. We first apply the peak detection algorithm to $(\mathbf{t}^0, \mathbf{M}^0)$ and obtain the peak locations $\mathcal{P}^0$. In Section 2.3.1 we use the marker peaks, a subset in $\mathcal{P}^0$ from reference lane 1s, to process $(\mathbf{t}^0, \mathbf{M}^0)$ into reference-aligned data $(\mathbf{t}^{\mathsf{R}}, \mathbf{M}^{\mathsf{R}})$. In Section 2.3.2, we transform the peaks detected from $(\mathbf{t}^{\mathsf{R}}, \mathbf{M}^{\mathsf{R}})$, denoted by $\mathcal{P}$, in a Bayesian framework to a joint posterior distribution of multivariate binary data $\mathbf{D}$ that represents presence or absence of a peak for all samples at a smaller number of landmarks, $L = 100$ in our application. In Section 3.2, we will process the reference-aligned high-frequency data $(\mathbf{t}^{\mathsf{R}}, \mathbf{M}^{\mathsf{R}})$ into $(\mathbf{t}, \mathbf{M})$ whose peaks are exactly matched to the landmarks present in $\mathbf{D}$.

2.2 *Peak Detection*

This section presents an algorithm for detecting the peaks $\mathcal{P}^0$ from standardized, high-frequency intensity data $(\mathbf{t}^0, \mathbf{M}^0)$. The peaks may appear with varying background intensities. Because the occurrence of a local maximum is thought to be more important than the background level in signature estimation, we design the algorithm to be insensitive to the absolute intensity.

We adopted the following peak detection algorithm:

i. *Local Difference Scoring.* Calculate the local difference score by comparing the intensity at bin b to its left and right neighbors exactly h bins away and to the local minimum intensity for bin $b = 1, \ldots, B$, lane $i = 1, \ldots, N_g$ of gel $g = 1, \ldots, G$. That is, we calculate

$$\begin{aligned}\mathsf{score}_{gi}(b) = \mathsf{sign}\left\{M^0_{gib} - M^0_{gi,\ell(b)}\right\} + \mathsf{sign}\left\{M^0_{gib} - M^0_{gi,r(b)}\right\} + \\ \mathsf{sign}\left\{M^0_{gib} - \min_{\ell(b)\le b'\le r(b)} M^0_{gib'} - C_0\right\},\end{aligned} \tag{2.1}$$

where $\mathsf{sign}(\mathsf{a}) = 1, 0, -1$ indicates positive, zero, or negative values; $\ell(b) = \max\{b - h, 1\}$

and $r(b) = \min\{b + h, B\}$ denote the left and right neighbors $h(= 10)$ bins away, and C_0 denotes the minimum peak relative elevation. The tunning parameter h controls the locality of the peaks and C_0 controls the minimum relative peak magnitude.

ii. *Peak Calling.* We look for the bins among peak candidates defined by $\{b \mid \mathsf{score}_{gi}(b) = \nu(= 3)\}$ that maximize their respective local intensities (see Appendix S2 for details and alternative peak calling methods). Let $\mathcal{P}^0_{gi}$ represent the collection of the peak locations for lane i and gel g.

Remark 1: Because the score defined in Step 3 depends on the intensity values only through local differences, the absolute background intensity and possible slowly changing baseline intensity are irrelevant. The local scoring method is fast and accurate. A 2-dimensional analogue has been used in astrophysics to find low grey-scale intensity galaxies from telescope images (Xu *and others*, 2016).

2.3 *Batch Effect Correction*

2.3.1 *Reference Alignment via Piecewise Linear Dewarping* Molecules with identical weight do not travel exactly the same distance along two arbitrary gels. Therefore, we first align the peaks on the reference lanes: $\mathcal{P}^0_{g1}$, $g = 1, \ldots, G$ via piecewise linear dewarping (Uchida and Sakoe, 2001). In our application, we used seven marker molecules of weight $(200, 116, 97, 66, 45, 31, 21.5)$ kDa.

We first exactly match the reference peaks $\mathcal{P}^0_{g1}$ on a query gel g to the reference peaks $\mathcal{P}_{g_0 1}$ on the template gel g_0, and then linearly stretch or compress the gels between the reference peaks. Quadratic or higher-order dewarping is also possible, but we found linear dewarping performs sufficiently well for our data. Appendix S3 gives details of the algorithm. We denote the high frequency, reference aligned data by $(\mathbf{t}^{\mathsf{R}}, \mathbf{M}^{\mathsf{R}}) = \{(t^{\mathsf{R}}_{gib}, M^{\mathsf{R}}_{gib})\}$; Let $\mathcal{P}$ collect the detected peaks.

2.3.2 *Correcting for Gel Deformation via Bayesian Image Dewarping* Another source of error during autoradiographic visualization is the non-rigid, spatial gel deformation. Middle panel of Figure 1(b) shows one such example. It also reveals three analytical challenges to be addressed before obtaining meaningful results from an automatic disease subsetting algorithm. First, some proteins, e.g., actin, are detected on multiple gels and must be aligned. The blue asterisks that denote the detected peaks near 0.43, form a smooth but non-linear curve from the top to the bottom of the gel. Second, fewer bands appear on the right half of the image, because these smaller proteins tend to contain fewer methionine residues for radiolabeling. Higher estimation uncertainty of the dewarping function is therefore expected for the right half. Third, the observed locations of the peaks are likely random around their true locations as the result of the multiple sources of error.

To address these issues, we designed a hierarchical Bayesian dewarping algorithm for two-dimensional images. The algorithm simulates presence/absence data from the conditional distribution of protein occurrence on a grid of equally-spaced landmark weights given the filtered $(\mathbf{t}^{\mathsf{R}}, \mathbf{M}^{\mathsf{R}})$ from the prior pre-processing. The stochastic model is defined on a coarser grid of landmark proteins, $\boldsymbol{\nu} = (0 = \nu_0 < \nu_1 < \ldots < \nu_L < \nu_{L+1} = 1)$ with $\nu_\ell = \ell/(L+1)$, $\ell = 0, 1, \ldots, L+1$. As defined further below, for each peak, the algorithm assigns a vector of probabilities to landmark proteins to optimize the joint probability of observing nearby peaks that drift across lanes. We introduce a novel shrinkage prior to promote alignment of peaks to a common landmark protein. We also introduce shrinkage priors that regularize the overall smoothness of the spatial dewarping functions.

Let (u_{gi}, T_{gij}) denote the (lane number, location) for peak $j = 1, \ldots, J_{gi}$ on lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$. We fix u_{gi} to take values in $\{1, 2, \ldots, N_g\}$ and collect them in $\boldsymbol{u} = \{u_{gij}\}$ where $u_{gij} = u_{gi}$ if they belong to the same lane i. Let $P_g = \sum_i J_{gi}$ denote the total number of peaks on gel g and $P = \sum P_g$. Let $\mathbf{T} = \{\boldsymbol{T}_g\}$, where $\boldsymbol{T}_g = (\ldots, T_{gi1}, T_{gi2}, \ldots, T_{gi,J_{gi}}, \ldots)'$ collects

the peak locations for gel $g = 1, \ldots, G$. Both $\boldsymbol{u}$ and $\boldsymbol{T}$ are P-dimensional column vectors. For computational stability, without changing notation, we standardize $\boldsymbol{u}$, $\boldsymbol{T}$ and $\boldsymbol{\nu}$ by substract their means and dividing by their standard deviations. We denote the data for the Bayesian dewarping model by $\mathcal{P} = \{\boldsymbol{u}, \boldsymbol{T}\}$ collect the locations of all the peaks.

Model Likelihood. *Peak-to-landmark indicators* $\mathbf{Z}$. Let Z_{gij} take values in $\{1, \ldots, L\}$. For example, $Z_{gij} = 3$ indicates that the j-th peak in lane i on gel g is aligned to Landmark 3. Note that any $\mathbf{Z}$ can be converted to multivariate binary data $\mathbf{D} = \{D_{gi\ell}\}$ for presence or absence of a detected peak at the landmarks by $D_{gi\ell} = \mathbf{1}\{\ell \in \{Z_{gij}, j = 1, \ldots, J_{gi}\}\}$, referred to as *signature*.

Gaussian mixture model for aligning observed peaks $\boldsymbol{T}$. We model $\boldsymbol{T} = \{T_{gij}\}$ as observations from a Gaussian mixture model with L components, each representing one landmark protein. Given $\boldsymbol{Z} = \{Z_{gij}\}$ and spatial dewarping function $\mathcal{S}_g$ to be discussed later, we assume

$$p\left\{ \left(\underbrace{u_{gi}}_{\substack{\text{lane}\\ \text{number}}}, \underbrace{T_{gij} = t}_{\substack{\text{peak}\\ \text{location}}} \right) \mid \underbrace{Z_{gij} = \ell}_{\substack{\text{matched to}\\ \text{landmark } \ell}}, \underbrace{T_{gi,j-1}}_{\substack{\text{nearest left}\\ \text{peak location}}}, \underbrace{\mathcal{S}_g}_{\substack{\text{warping}\\ \text{function}}}, \underbrace{\sigma_\epsilon}_{\substack{\text{noise}\\ \text{level}}} \right\} = \begin{cases} \phi(t; \mathcal{S}_g(u_{gi}, \nu_\ell), \sigma_\epsilon), & t \in \mathcal{I}_{gij}(\nu_\ell, A_0); \\ 0, & \text{otherwise,} \end{cases} \tag{2.2}$$

$\ell = 1, \ldots, L$, for peak $j = 1, \ldots, J_{gi}$, lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$, where $\phi(\cdot; a, b)$ is the Gaussian density function with mean a and standard deviation b, and $\mathcal{S}_g(u, \nu)$ is an unknown smooth bivariate function that characterizes the spatial gel deformation $(u_{gi}, \nu_\ell) \mapsto (u_{gi}, \mathcal{S}_g(u_i, \nu_\ell))$.

Remark 2: The peak location T_{gij} follows Gaussian distribution with mean equal to ν_ℓ plus a horizontal displacement $\mathcal{S}_g(u_{gi}, \nu_\ell)$ and noise level σ_ϵ. We assume σ_ϵ is independent of landmark and lane. The density function (2.2) is supported in the set $\mathcal{I}_{gij}(\nu_\ell, A_0) \triangleq \{t : |t - \nu_\ell| < A_0 \text{ and } t > T_{gi,j-1}\}$. The first inequality prohibits T_{gij} being matched to distant landmarks and the second prevents reverse warping, i.e., ensures $Z_{gij} \leq Z_{gij'}$ whenever $T_{gij} \leq T_{gij'}$.

Bivariate smooth warping functions $\mathcal{S}_g$. For gel g, we model the warping function $\mathcal{S}_g : \mathbb{R}^2 \to \mathbb{R}$

that warps the landmarks (u_{gi}, ν_ℓ) horizontally to $(u_{gi}, \mathcal{S}_g(u_{gi}, \nu_\ell))$ using tensor product basis expansion

$$\mathcal{S}_g(u,\nu) = \sum_{s=1}^{T_u}\sum_{t=1}^{T_\nu} \beta_{gst} B_{g1s}(u) B_{g2t}(\nu), \tag{2.3}$$

where $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$ are the s-th and t-th cubic B-spline basis with intercept anchored at knots $\boldsymbol{\kappa}_u$ and $\boldsymbol{\kappa}_\nu$ along the two coordinate directions, respectively (Friedman *and others*, 2001, Chapter 5) and $T_u = |\boldsymbol{\kappa}_u|$, $T_\nu = |\boldsymbol{\kappa}_\nu|$ are the total number of basis functions in $u-$ and $\nu-$ direction. In subsequent analyses, we choose $\boldsymbol{\kappa}_\nu$ with $T_\nu - 4$ internal knots at the $s/(T_\nu - 3)$-th quantile of $\{T_{gij}\}$, $s = 1, \ldots, T_\nu - 4$ and similarly for $\boldsymbol{\kappa}_u$.

However, valid spatial gel deformations are gel stretching, compression or shift along the ν direction. We thus constrain the shape of $\mathcal{S}_g, g = 1, \ldots, G$ by

$$\textit{Monotonicity}: \ \nu_0 < \mathcal{S}_g(u, \nu_{\ell-1}) < \mathcal{S}_g(u, \nu_\ell) < \nu_{L+1}, \forall \ell = 1, \ldots, L, \forall u; \tag{2.4}$$

$$\textit{Boundary Constraint}: \ \mathcal{S}_g(u, \nu_0) = \nu_0, \mathcal{S}_g(u, \nu_{L+1}) = \nu_{L+1}. \tag{2.5}$$

The first constraint prevents reverse gel dewarping and the second assumes away gel shifting; it can be relaxed to allow horizontal shifts by adding/substracting Δ for both equalities. We implement both constraints by requiring the B-spline coefficients $\boldsymbol{\beta}_g = \{\beta_{gst}\}_{s=1,t=1}^{T_u, T_\nu}$ to satisfy: $\nu_0 = \beta_{gs1} < \beta_{gs2} < \ldots < \beta_{gs,T_\nu - 1} < \beta_{gsT_\nu} = \nu_{L+1}$, $\forall s = 1, \ldots, T_u$. Although only sufficient for $\mathcal{S}_g$'s monotonicity and boundary constraints, the $\boldsymbol{\beta}_g$ constraints allow flexible and realistic warpings. Figure 3 shows a member warping function that corrects for local L-, S- and 7-shaped deformations. This approach extends the curve registration method (Telesca and Inoue, 2008) to two-dimensional surfaces without the self-similarity assumption.

Priors. *Prior for* $\boldsymbol{Z}$. We describe a shrinkage prior for $\mathbf{Z}$ motivated by the needs 1) to align the actin peaks (as in middle panel of Figure 1(b)) to an identical landmark in a single gel, and 2) to share the information about the location of actin peaks across multiple gels. We specify the prior distribution based on non-homogeneous Poisson processes with extreme intensities at

a small number of landmarks. For each landmark, the intensity is further shared across multiple gels to facilitate borrowing of information.

Let the total number of observed peaks follow a Poisson distribution: $J_{gi} \overset{d}{\sim} \mathsf{Poisson}(\Lambda_g)$, for lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$. Given J_{gi}, let $Z^*_{gij} \overset{iid}{\sim} \mathsf{Categorical}\left(\{\lambda^*_\ell\}_{\ell=1}^L\right)$ describe which landmarks are present in lane i of gel g. Because $\{Z_{gij}, j = 1, \ldots, J_{gi}\}$ are by definition ordered on each lane, we increasingly sort $\{Z^*_{gij}\}$. For hyperpriors, let $\lambda^*_\ell = \lambda_\ell / \sum_{\ell'} \lambda_{\ell'}$ where $\lambda_\ell \overset{iid}{\sim} \mathsf{Normal}(0, \tau), \ell = 1, \ldots, L$, and the hyperparameter $\tau \overset{d}{\sim} \mathsf{Inv\text{-}Gamma}(10^{-4}, 10^{-4})$. Integrating over τ, λ_ℓ is t-distributed.

Remark 3: The intensity parameters $\{\lambda^*_\ell\}$, one per landmark, *in a priori* determines the probability of a landmark protein present on one lane: $\mathbb{P}(D_{gi\ell} = 1 \mid \lambda^*_\ell) \approx 1 - \exp(-\lambda^*_\ell)$ if L is large. By (A2), the ratio of conditional posterior probabilities of assigning the peak T_{gij} to one versus another landmark is $\frac{\phi(T_{gij}; \mathcal{S}_g(u_{gi}, \nu_\ell), \sigma)}{\phi(T_{gij}; \mathcal{S}_g(u_{gi}, \nu_{\ell'}), \sigma)} \cdot \frac{1-\exp(-\lambda^*_\ell)}{1-\exp(-\lambda^*_{\ell'})}$. Suppose $\lambda^*_\ell > \lambda^*_{\ell'}$, the second ratio favors ν_ℓ, unless the likelihood ratio in the first term is small. The $\{\lambda^*_\ell\}$ are independent of g and i hence globally modulate the probability of landmark presence for all the gels. For subsequent analyses, we withhold prior biological knowledge about the prevalent landmark proteins, and instead assign independent t-distributed prior for λ_ℓs, whose heavy tails generate occasional large λ^*_ℓ values. The posterior inference algorithm will occasionally visit the $\mathbf{Z}$-configuration that, say many $Z_{gij} = \ell$, which if increases the joint posterior, will retain such configuration and identify important landmark ℓ.

Prior for $\boldsymbol{\beta}_g$. We first specify priors for the horizontal basis coefficients β_{gst}, $t = 2, \ldots, T_\nu - 1$ at the u-direction basis $s = 1$. We use a first-order random walk prior (Lang and Brezger, 2004)

$$\beta_{gst} - \beta^{\mathsf{id}}_{t-1} \overset{d}{\sim} N(\cdot; \beta_{gs,t-1} - \beta^{\mathsf{id}}_t, \sigma^2_{g1}) I(\beta_{gs,t-1}, \nu_{L+1}), s = 1, t = 2, \ldots, T_\nu - 1, \tag{2.6}$$

where $\boldsymbol{\beta}^{\mathsf{id}} = (\beta^{\mathsf{id}}_1, \ldots, \beta^{\mathsf{id}}_{T_\nu})'$ is the vector of coefficients for identity warping function $S(u, \nu) = \nu$. The hyperparameter σ^2_{g1} controls the similarity of $\{\beta_{1t}\}_{t=1}^{T_\nu}$ to $\boldsymbol{\beta}^{\mathsf{id}}$ and hence the similarity of

$\mathcal{S}(u_2, \nu)$ to identify function; $\sigma_{g1}^2 = 0$ represents no warping. We refer σ_{g1}^{-2} as the smoothing parameter in the ν-direction.

Next, for any $t = 2, \ldots, T_\nu - 1$, we specify another random walk prior for the vertical basis coefficients β_{gst}, $s = 1, \ldots, T_u$:

$$\beta_{gst} \overset{d}{\sim} N(\cdot; \beta_{g,s-1,t}, \sigma_{gt}^2). \tag{2.7}$$

Similarly, the hyperparameter $\{\sigma_{gt}^2\}$ is the smoothness parameter of for $\mathcal{S}_g$ in the vertical $u-$direction; $\sigma_{gt}^2 = 0$ means identical amount of warping across lanes. Details about the hyperpriors on σ_{g1}^2 and $\{\sigma_{gt}^2\}$ are provided in Appendix S5.

Joint Distribution. Let $\boldsymbol{B}_g$ be a matrix with P rows, each defined by $\boldsymbol{B}_{g1}(u_{gi})' \otimes \boldsymbol{B}_{g2}(\nu_{Z_{gij}})'$ for a peak (u_{gi}, T_{gij}), where $\nu_{Z_{gij}}$ is the aligned landmark, $\boldsymbol{B}_{g1}(u) = (B_{g11}(u), \ldots, B_{g1T_u}(u))'$ and $\boldsymbol{B}_{g2}(\nu) = (B_{g21}(\nu), \ldots, B_{g2T_\nu}(\nu))'$ are the B-spline bases evaluated at u and ν, respectively. Let $\mathsf{vec}(\boldsymbol{\beta}_g')$ be a column vector that stacks the rows of $\boldsymbol{\beta}_g$. We obtain the join distribution

$$p(\boldsymbol{\lambda}^*) \times \prod_{g=1}^{G} \Bigg\{ p(\sigma_{g1}^2) \prod_{t=2}^{T_\nu - 1} p(\sigma_{gt}^2, \rho_g) \times N_P\left(\boldsymbol{T}_g; \boldsymbol{B}_g \mathsf{vec}[\boldsymbol{\beta}_g'], \sigma_\epsilon^2 \mathbf{I}\right)$$
$$\times N_{T_\nu - 1}\left(\{\beta_{g1t}\}_{t=1}^{T_\nu - 1}; \boldsymbol{\beta}_{[-T_\nu]}^{id}, \sigma_{g1}^2 \mathbf{I}\right) \times \prod_{t=2}^{T_\nu - 1} N_{T_u}\left(\{\beta_{gst}\}_{s=1}^{T_u}; \mathbf{0}, \sigma_{gt}^2 \mathbf{I}\right) \prod_{ij} \mathsf{OrderedCategorical}\left(Z_{gij}; \boldsymbol{\lambda}_\ell^*\right) \Bigg\}, \tag{2.8}$$

where $p(\boldsymbol{\lambda}^*)$, $p(\sigma_{g1}^2)$ and $p(\sigma_{gt}^2, \rho_g)$ are the priors and hyperpriors and $N_d(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the d-dimensional multivariate normal density with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

3. Model Estimation and Implementation

3.1 *Posterior Sampling*

We use Markov chain Monte Carlo (MCMC) for versatile posterior inference by simulating samples from the joint posterior distributions of all the unknowns (e.g., Gelfand and Smith, 1990). Based on the samples obtained from MCMC, we can perform posterior inferences about any

functionals of model parameters, e.g., the gel warping functions $\{\mathcal{S}_g(\boldsymbol{\beta})\}$, the peak-to-landmark alignment indicators $\mathbf{Z}$. Appendix S4 provides full details about the sampling algorithm along with discussions of *scattering condition* that ensures statistical identifiability of the warping functions. Subsequent posterior analyses were based on three chains of 10,000 iterations with a burn-in period of 5,000 iterations. We monitor the convergence by chain histories, auto-correlations, kernel density plots, and Brooks-Gelman-Rubin statistic. Convergence is fast within thousands of burn-in iterations. All model estimation and visualization is performed by the `R` package `spotgear` (https://github.com/zhenkewu/spotgear).

Turning to dewarping a new GEA image, let $\mathcal{D}_{g^*}$ be new raw intensity data after reference alignment described in Section 2.3.1. We can approximate the joint posterior of unknown basis coefficients and peak-to-landmark indicators by

$$p(\boldsymbol{\beta}_{g^*}, \boldsymbol{Z}_{g^*} \mid \mathcal{D}, \mathcal{D}_{g^*}) = \int p(\boldsymbol{\beta}_{g^*}, \boldsymbol{Z}_{g^*} \mid \boldsymbol{\lambda}, \mathcal{D}_{g^*}) p(\boldsymbol{\lambda} \mid \mathcal{D}, \mathcal{D}_{g^*}) \mathrm{d}\boldsymbol{\lambda} \approx \int p(\boldsymbol{\beta}_{g^*}, \boldsymbol{Z}_{g^*} \mid \boldsymbol{\lambda}, \mathcal{D}_{g^*}) p(\boldsymbol{\lambda} \mid \mathcal{D}) \mathrm{d}\boldsymbol{\lambda},$$

where the two terms in the integrand are the one-sample conditional posterior and the posterior of $\boldsymbol{\lambda}$ given the preprocessed data $\mathcal{D}$. The first term can be calculated from the joint distribution (2.8) and the integral is readily estimated by $\sum_t p(\boldsymbol{\beta}_{g^*}, \boldsymbol{Z}_{g^*} \mid \boldsymbol{\lambda}^{(t)}, \mathcal{D}_{g^*})$ using the stored posterior samples $\{\boldsymbol{\lambda}^{(t)}\}$.

3.2 *Approximation for Dewarping Function $\mathcal{S}_g$*

We also need to produce exact peak-aligned high-frequency data $(\mathbf{t}, \mathbf{M})$ for disease subsetting in Section 4. However, because $\mathcal{S}_g$ is the mean surface and does not account for the noise introduced by σ_ϵ^2, reverse mapping of $\mathcal{S}_g$ from $(\mathbf{t}^{\mathsf{R}}, \mathbf{M}^{\mathsf{R}})$ cannot do exact peak alignment. Instead, for each sample lane, because $\{Z_{gij}\}_j$ maps the peaks to the corresponding landmark proteins, as an approximation, we simply perform piecewise linear dewarping of reference aligned high-frequency data $(\mathbf{t}^{\mathsf{R}}, \mathbf{M}^{\mathsf{R}})$ anchoring at the landmarks $\{Z_{gij}\}_j$ and the two endpoint landmarks ν_0 and ν_{L+1}. One can view the Bayesian dewarping model first estimates $\mathcal{S}_g$ based on a sparse grid of

landmarks to encourage nearby peaks to be aligned. We then can ignore the $\mathcal{S}_g$ estimates and use the estimated peak-to-landmark indicators $\boldsymbol{Z}$ for exact peak matching. In subsequent analyses, we use the *maximum a posteriori* (MAP) $\{\hat{Z}_{gij}\}$ to construct this approximation.

4. Application to Scleroderma

We used sera from well-characterized patients with scleroderma and an associated cancer identified through the IRB-approved Johns Hopkins Scleroderma Center database (Shah *and others*, 2017). We first analyze data comprised of GEA replicates on 20 samples and thus 20 experimental pairs of size two. Compared to hierarchical clustering without preprocessing, we show the preprocessing creates better separated clusters and hence improves the accuracy of cluster detection when compared to the truth defined by the replication. Based on a second set of GEA data without replicates we apply the preprocessing method and identify strong clusters that are well-separated and scientifically meaningful.

4.1 *Outline of Analyses*

Firstly, we preprocess the raw images. We apply the peak detection algorithm in Section 2.2 followed by batch effect corrections as described in Section 2.3. We exclude reference lane 1s for 2D Bayesian dewarping. We used $T_u = 6$ and $T_\nu = 10$ cubic B-spline basis functions in the vertical and horizontal directions, respectively. The 2D smooth dewarping functions for all the gels are then estimated via the posterior mean dewarpings $\{\hat{\mathcal{S}}_g = \mathcal{S}_g(\cdot, \cdot; \hat{\boldsymbol{\beta}}_g)\}$, where $\hat{\boldsymbol{\beta}}_g$ is the posterior mean estimated by the empirical average of the MCMC samples. We also obtain the *maximum a posteriori* peak-to-landmark indicators $\hat{\boldsymbol{Z}} = \{\hat{Z}_{gij}\}$.

On the other hand, for every sample lane, as described in Section 3.2, we perform *exact* matching of the observed peaks of identical weights. Based on the exact peak-aligned images $\mathbf{M}$, for a pair of sample lanes i and i', we calculate the pairwise distances $d(i, i') = 1 - \widehat{\mathsf{cor}}(\mathbf{M}_{gi\cdot}, \mathbf{M}_{gi'\cdot})$

where $\mathbf{M}_{gi\cdot} = (M_{gi1}, \ldots, M_{giB})'$ and $\widehat{\mathsf{cor}}(\cdot,\cdot)$ is the Pearson's correlation coefficient. We use the obtained distance matrix $\hat{D}$ for agglomerative hierarchical clustering with complete linkage to produce a dendrogram $\hat{\mathcal{T}} = \mathcal{T}(\hat{D})$. Let $\hat{\mathcal{C}}(n)$, $n = 1, \ldots, N$ represent a nested set of clusters depending on where the dendrogram is cut. We similarly denote the dendrogram produced without preprocessing by $\hat{\mathcal{T}}^0 = \mathcal{T}(D^0)$ and the nested clusters by $\{\hat{\mathcal{C}}^0(n)\}$, respectively.

When the true clustering is given, for example, in replication experiments, we will assess the agreement of $\mathcal{C}(n)$ and $\mathcal{C}^0(n)$ with the truth $\mathcal{C}^*$, for the number of clusters $n = 2, \ldots, N/2$. Adjusted Rand index (aRI; Hubert and Arabie (1985)) can assess the similarity of two ways of partitioning the same set of observations and can handle different numbers of clusters. ARI is defined by

$$\mathsf{aRI}(\mathcal{C}, \mathcal{C}') = \frac{\sum_{r,c} \binom{n_{rc}}{2} - \left[\sum_r \binom{n_{r\cdot}}{2} \sum_c \binom{n_{\cdot c}}{2}\right] / \binom{N}{2}}{0.5 \left[\sum_r \binom{n_{r\cdot}}{2} + \sum_c \binom{n_{\cdot c}}{2}\right] - \left[\sum_r \binom{n_{r\cdot}}{2} \sum_c \binom{n_{\cdot c}}{2}\right] / \binom{N}{2}}, \tag{4.9}$$

where n_{rc} represents the number of observations placed in the rth cluster of the first partition $\mathcal{C}$ and in the cth cluster of the second partition $\mathcal{C}'$, $\sum_{r,c} \binom{n_{rc}}{2} (\leq 0.5 \left[\sum_r \binom{n_{r\cdot}}{2} + \sum_c \binom{n_{\cdot c}}{2}\right])$ is the number of observation pairs placed in the same cluster in both partitions and $\sum_r \binom{n_{r\cdot}}{2}$ and $\sum_c \binom{n_{\cdot c}}{2}$ calculate the number of pairs placed in the same cluster for the first and the same cluster for second partition, respectively. ARI is bounded between -1 and 1 and corrects for chance agreement: it equals one for identical clusterings and is on average zero for two random partitions with larger values indicating good agreements.

We also evaluate the clustering strength of $\hat{\mathcal{C}}(n)$ and $\hat{\mathcal{C}}^0(n)$, for $n = 2, \ldots, N/2$, by calculating the average silhouette based on the comparison between cluster tightness and separation (Rousseeuw, 1987). For observation i, its silhouette $s(i)$ with respect to an arbitrary partition $\mathcal{C}$ compares the within- to between-cluster average distances: $s(i) = [b(i) - a(i)]/\max\{a(i), b(i)\}$ where $a(i)$ is the average distance of i to all other observations within the same cluster and $b(i) = \min_{C \in \mathcal{C}: i \notin C} \frac{\sum_{i' \in C} d(i,i')}{|C|}$ is the minimum average distance between i and a cluster not containing i. $s(i)$ lies in $[-1, 1]$ with a large value indicating observation i in a tight and isolated

levels for phylogenetic trees. *Proceedings of the National Academy of Sciences* **93**(23), 13429–13429.

Friedman, Jerome, Hastie, Trevor and Tibshirani, Robert. (2001). *The elements of statistical learning*, Volume 1. Springer Series in Statistics Springer, Berlin.

Gelfand, Alan E and Smith, Adrian FM. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* **85**(410), 398–409.

Hubert, Lawrence and Arabie, Phipps. (1985). Comparing partitions. *Journal of classification* **2**(1), 193–218.

Lang, Stefan and Brezger, Andreas. (2004). Bayesian p-splines. *Journal of computational and graphical statistics* **13**(1), 183–212.

Miller, Jeffrey W and Harrison, Matthew T. (2015). Mixture models with a prior on the number of components. *arXiv preprint arXiv:1502.06241*.

Rosen, Antony and Casciola-Rosen, Livia. (2016). Autoantigens as partners in initiation and propagation of autoimmune rheumatic diseases. *Annual review of immunology* **34**, 395–420.

Rousseeuw, Peter J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65.

Shah, Ami A., Xu, George, Rosen, Antony, Hummers, Laura K., Wigley, Fredrick M., Elledge, Stephen J. and Casciola-Rosen, Livia. (2017). Anti-rnpc3 antibodies as a marker of cancer-associated scleroderma. *Arthritis & Rheumatology*, n/a–n/a.

Shimodaira, Hidetoshi. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic biology* **51**(3), 492–508.

Shimodaira, Hidetoshi *and others*. (2004). Approximately unbiased tests of regions using

cluster. A large average silhouette $N^{-1}\sum s(i)$ indicates well separated and tight clusterings.

The structural uncertainty of the finite-sample dendrogram estimate $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}}^0$ can be assessed via multiscale bootstrap resampling (Shimodaira *and others*, 2004). The multiscale bootstrap resampling (MBR) is a method to perturb the data and assess the confidence levels for the presence of each subtree in the estimated dendrogram (e.g., Shimodaira, 2002; Efron *and others*, 1996). MBR calculates the frequency with which a subtree appears in an estimated dendrogram across all bootstrap iterations. A bias-corrected frequency, referred to as the approximately unbiased (AU) probability value will be used to determine the strength of evidence for the presence of a subtree, where a large value (e.g., > 0.95) indicates strong evidence.

4.2 *Replication Experiments*

Twenty samples are tested each with two different lengths of exposure to autoradiography devices, long (two-week) versus short (one-week). We thus obtain 40 lanes on two gel images that form 20 replicate pairs. Each gel image is has 19 serum sample lanes plus one reference sample lane comprised of marker molecules with known molecular weights. The posterior dewarping results by the 2D Bayesian dewarping are shown in Appendix Figure S2.

Upon hierarchical clustering, we assess the agreement of the estimated clusterings $\hat{\mathcal{C}}(n)$ and $\hat{\mathcal{C}}^0(n)$ with the true replication-based clusters for the number of clusters $n = 2, \ldots, 20$. For every n, we calculate the adjusted Rand Index (aRI) and obtain its confidence intervals by bootstrapping. Specifically, to account for the inherent replication design, we repeat for $b = 1, \ldots, B = 1000$ the following procedure: 1) resample replicate pairs with replacement, 2) calculate the boostrapped distance matrix $\hat{D}^{*(b)}$, 3) obtain dendrogram by hierarchical clustering, 4) cut the dendrogram at various levels to form clusters and compare them to the truth. Note that the truth might contain clusters of size four or larger doubles of two because one pair can be resampled more than once.

Figure 4 shows that preprocessing achieves overall higher mean aRI, $\mathsf{aRI}(\hat{\mathcal{C}}(n), \mathsf{True\ Pairs})$,

across different numbers of clusters n. One of the benefits is that we improved the ability of the hierarchical clustering to detect the true replicate pairs. Compared to the analysis without preprocessing, we identified more replicate pairs on terminal leaves (13 versus 8 out of the 20 true replicate pairs) with percent reductions in the within-pair distance ranging from $6.2-66.4\%$ (mean 26.9%). For example, if 20 clusters are formed by dendrogram cutting, the aRI is 0.69 (95% confidence interval: (0.51,0.89)) compared to 0.47 $(0.30, 0.68)$ for the analysis without preprocessing. The $\mathsf{aRI}(\hat{\mathcal{C}}(n), \mathsf{True\ Pairs})$ and $\mathsf{aRI}(\hat{\mathcal{C}}^0(n), \mathsf{True\ Pairs})$ are most discrepant at $n = 18$: 0.73 $(0.52, 0.91)$ with preprocessing versus 0.49 $(0.29, 0.73)$ otherwise.

Confidence levels of the presence of true replicate pairs are also much improved by preprocessing. Appendix Figure S3 examines the confidence levels associated with each subtree (numbered edges) with and without preprocessing. There are uniform increases in the confidence levels for many clusters defined by the subtrees. For example, for pair 18, the confidence level increases from 0.67 to 0.99 after preprocessing; The confidence levels for detecting the pairs 2, 8 and 11 see similar increases from (0.68, 0.59, 0.67) to (0.96, 0.92, 0.93) after preprocessing, as confirmed by the $14.2-117.6\%$ percent increase, $0.03-0.18$ absolute increase in cluster separation as measured by average silhouette. The better separation and tighter clusters provided by preprocessing also lead to more parsimonious clusters, e.g., the dendrogram with preprocessing correctly excluded the subtree 35 and 37 in $\hat{\mathcal{T}}^0$ at the bottom of Appendix Figure S3.

4.3 *Scleroderma GEA Data without Replicates*

We conducted GEA on four sets of sera from scleroderma patients with cancer who are all negative for autoantibodies to RNA polymerase III, topoisomerase I and centromere proteins. The status of any other specificities in these sera, whether defined or novel, was not known at the time this study was done. Each gel is loaded with 19 serum samples (loaded in random order in the gel lanes) and one reference sample comprised of molecules with known molecular weights. Results

of their joint analysis are discussed below.

We applied the preprocessing methods described above to the four gel sets. We first removed a few spots on the right of the gels caused by localized gel contamination assuming absence of peaks there. The posterior dewarping results are shown in Figure 5. Each detected peak $\{T_{gij}\}$ shown by a blue dot is connected to a red triangle that represents the landmark $\hat{Z}_{gij}$ that maximizes the marginal posterior probability: $\hat{Z}_{gij} = \arg\max_\ell \mathbb{P}(Z_{gij} = \ell \mid \mathcal{P}_{gi}, g = 1, 2, 3, 4, i = 2, \ldots, 20)$. The vertical bundle of black curves, one per landmark, shows the global shape of the estimated warping functions $\hat{\mathcal{S}}_g$, $g = 1, 2, 3, 4$, where $\hat{\mathcal{S}}_g = \mathcal{S}_g(\cdot, \cdot; \hat{\boldsymbol{\beta}}_g)$ and $\hat{\boldsymbol{\beta}}_g$ is the posterior mean. The locations traced by the same curve are estimated to represent identical molecular weights. Bottom of Figure 5 shows the marginal posterior probabilites of each landmark being matched with a peak for one sample. For example, the posterior probability is 0.59 for Landmark 50 ($\sim$ 43.4 kDa): MAP of $\hat{\boldsymbol{Z}}$ shows that 73 out of 76 lanes have a peak being matched to it. In gel 1, this high probability caused the many detected peaks (blue dots) to the right of 45 kDa marker aligned altogether to Landmark 50. Note that we did not use any prior knowledge that actin is present in all samples here. The marginal posterior probability is expected to further increase when more samples containing actin are combined for hierarchical Bayesian dewarping. Landmark 46 ($\sim$ 46.6 kDa) is another molecular hotspot where 54 out of 76 lanes have matched peaks. On the other hand, for example, 18 and 1 out of 76 are matched to Landmarks 36 ($\sim$ 59.8 kDa) and 89 ($\sim$ 23.4 kDa), respectively. Their marginal posterior probabilities are hence low at 0.21 and 0.01.

An animation of the continuous dewarping process is available at https://github.com/zhenkewu/spotgear. It matches the detected peaks T_{gij} to their inferred landmarks $\hat{Z}_{gij}$ and morphs the posterior mean dewarping $\hat{\mathcal{S}}_g$ into constant function $\mathcal{I} : (u, \nu) \mapsto (u, \nu)$. There also shows the preprocessed high-frequency data with exactly matched peaks as described in Section 3.2.

Preprocessing eliminates many huge clusters that are otherwise formed without preprocessing as shown at the bottom of Appendix Figure S4. The percent and absolute increases in the

average silhouette are between $8.8 - 39.5\%$ and $0.02 - 0.08$ respectively for varying n upon prepreocessing. The better separation enabled by the proposed preprocessing corrected potential misaligned cross-lane, removed global warping phenomena and revealed a few strong clusters after maximal separation observations. The clusters with 0.95 confidence levels or higher are shown in red rectangles in Appendix Figure S4 for the analyses done with preprocessing (top) and without preprocessing (bottom). For the former, the first cluster from the right (number 44) comprises of seven sample lanes ((Set, Lane): (1,19), (4,3), (1,18), (3,8), (4,10), (2,4), (2,13)) enriched at $\sim$ 32.7 and $\sim$ 27.9 kDa which is split into two clusters (number 47 and 14) for the analyses without preprocessing. Enriched at $\sim$ 103.4 kDa, Clusters 46 at the bottom and 40 at the top are comprised of identical samples with improvement in the confidence level from 0.97 to 1 after preprocessing.

4.4 *Additional Validation of the Algorithm Method*

We selected one prominent cluster to validate the method. Before knowing the algorithm clustering results, an experienced investigator carefully reviewed the 4 sets of immunoprecipitation data and assigned groups based on visual band patterns and sizes. Four sera were assigned to a group based on a pattern consistent with antibodies against a known autoantigen (termed anti-PMScL). The algorithm identified these same 4 sera as a cluster (lanes marked in red font on Appendix Figure S4). All 76 sera were tested using a commercially available line immunoblot assay (EuroImmun; Systemic Sclerosis (Nucleoli) profile) to determine which of the sera had antibodies against PMScL. Only 4 sera were positive for this antibody specificity, and they were identical to those assigned to this cluster by both the investigator and the algorithm.

It is noteworthy that the algorithm detected several other clusters of > 3 sera, e.g., Number 34 ((Set, Lane): (3,14), (3,12), (4,4)), 38 (Reference Lanes), 50 ((2,11), (3,18), (2,9), (2,12)) and 53 ((1,7), (1,20), (2,16)). After reviewing the clusters identified by the algorithm, the experienced

investigator went back to the original immunoprecipitation data again and confirmed that the antibody patterns of these sera were indeed similar enough to cluster, and warrant further investigation for discovery of the relevant specificities. Because the dendrogram is constructed using the distance matrix $\hat{D}$ that is informed by the preprocessing step, observations placed together in the subtrees, although not detected with $> 95\%$ confidence level, can guide subsetting and validation.

5. Discussion

In this article, we have developed a novel statistical approach to preprocessing and analyzing two-dimensional image data obtained from gel electrophoresis autoradiography with the objective of detecting autoimmune disease subsets based on autoantibody signatures. The hierarchical Bayesian image dewarping model provides a natural framework for assessing uncertainty in the estimated alignment and warping functions and through MCMC sampling technique allows us to derive inferences about a richer set of quantities of interest.

Through the analyses of two sets of gel data from scleroderma patients, with and without replicates, we have shown that the sample lanes are better compared and clusters are better separated and more accurately estimated upon hierarchical Bayesian dewarping. We also studied the performance of naive analysis without the proposed preprocessing. We conclude that there is added benefits of the proposed automated procedure to estimating disease subsets compared to the naive analysis and human recognition of band patterns scattered across multiple gels, and hence provides a useful improvement for researchers using gel electrophoresis to study differential autoantibody compositions among disease subgroups. We expect marginal though worthwhile gains to be achievable by using more carefully designed and tested tuning parameter selection procedure for local scoring (Section 2.2).

Multiple extensions to the proposed method that build on biological processes warrant fu-

ture research. First, in our hierarchical Bayesian dewarping model, we assume that the intensity parameters $\{\lambda_\ell^*\}$ of alignment to each landmark are drawn from a common set of population distributions. Autoantibody presence or absence may differ across samples, however. For example, cancer versus non-cancer patients may have distinct priors of certain autoantibody presence/absence. We can either add another hierarchy for Bayesian dewarping or develop regression models for $\{\lambda_\ell^*\}$ to incorporate disease phenotype information or other covariates, e.g, age and gender to refine disease subsetting. Second, multiple autoantibodies produced against a particular molecular complex are considered to be present or absent in a grouped fashion. This intermolecular spreading of the immune response to multiple components linked within a multimolecular complex is an important property of the immune response, reflecting the ability of B cells to use their specific surface immunoglobulin to capture whole molecular complexes through binding to the single component that they recognize, and then driving additional immune response to other components of the complex. The biological structure can be represented by a binary matrix $\mathbf{M}_{C\times L}$, one row per complex, where $\{M_{c\ell}\}_{\ell=1}^{L}$ is a multivariate binary vector with 1 for presence of landmark ℓ in complex c and 0 otherwise. The complexes are then assembled via $\boldsymbol{\eta}_{N\times L} = \mathbf{AM}$ to produce the actual presence or absence of landmarks for every patient, where $\mathbf{A}$ is a $N \times C$ binary matrix with one assembly vector per row representing presence or absence of the list of complexes. Prior biological knowledge can be readily implemented via constraints on $\mathbf{A}$ or $\mathbf{M}$. For example, $A_{i1} = 1$ for all subjects acknowledges the universal presence of autoantibodies produced by Complex 1, e.g., actin and likely others. $\mathbf{A}$ and $\mathbf{M}$ can be inferred from alignment indicators $\mathbf{Z}$ or extracted continuous intensity shape information for each landmark and lane either by regularization or using shrinkage priors in a Bayesian framework for encouraging few and maximally different complexes (e.g., Broderick *and others*, 2013; Miller and Harrison, 2015). Our preliminary results (not shown here) show good subset and signature estimation performance. One practical advantage of the Bayesian complex assembly approach lies in its convenient accommodation of

repeated GEA on the same unknown sample by equality constraints on rows of **A**. Models for repeated autoantibody measurements across multiple clinic visits are also important. Finally, the latent variable formulation of the dewarping enables easy coupling with general latent variable models with discrete state space and factorization structures that incorporate multiple sources of lab test and phenotype data, facilitate definition of disease subgroups and perform individual predictions (e.g., Coley *and others*, 2016; Wu *and others*, 2016, 2017).

Supplementary Materials

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

Research reported in this work was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1408-20318) and a generous grant from the Jerome L. Greene Foundation. The Johns Hopkins Rheumatic Disease Research Core Center, where the sera were processed and banked, and the antibody assays were performed, is supported by NIH grant P30 AR-070254.

References

Billera, Louis J, Holmes, Susan P and Vogtmann, Karen. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics* **27**(4), 733–767.

Broderick, Tamara, Jordan, Michael I. and Pitman, Jim. (2013, 08). Cluster and feature modeling from combinatorial stochastic processes. *Statist. Sci.* **28**(3), 289–312.

Coley, Rebecca Yates, Fisher, Aaron J., Mamawala, Mufaddal, Carter, Herbert Ballentine, Pienta, Kenneth J. and Zeger, Scott L. (2016). A bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer. *Biometrics*, In Press.

Efron, Bradley, Halloran, Elizabeth and Holmes, Susan. (1996). Bootstrap confidence

multistep-multiscale bootstrap resampling. *The Annals of Statistics* **32**(6), 2616–2641.

Telesca, Donatello and Inoue, Lurdes Y T. (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association* **103**(481), 328–339.

Uchida, Seiichi and Sakoe, Hiroaki. (2001). Piecewise linear two-dimensional warping. *Systems and Computers in Japan* **32**(12), 1–9.

Willis, A. and Bell, R. C. (2016, November). Uncertainty in phylogenetic tree estimates. *ArXiv e-prints*.

Wu, Zhenke, Deloria-Knoll, Maria, Hammitt, Laura L and Zeger, Scott L. (2016). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(1), 97–114.

Wu, Zhenke, Deloria-Knoll, Maria and Zeger, Scott L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics* **18**(2), 200.

Xu, Bingxiao, Postman, Marc, Meneghetti, Massimo, Seitz, Stella, Zitrin, Adi, Merten, Julian, Maoz, Dani, Frye, Brenda, Umetsu, Keiichi, Zheng, Wei *and others*. (2016). The detection and statistics of giant arcs behind clash clusters. *The Astrophysical Journal* **817**(2), 85.

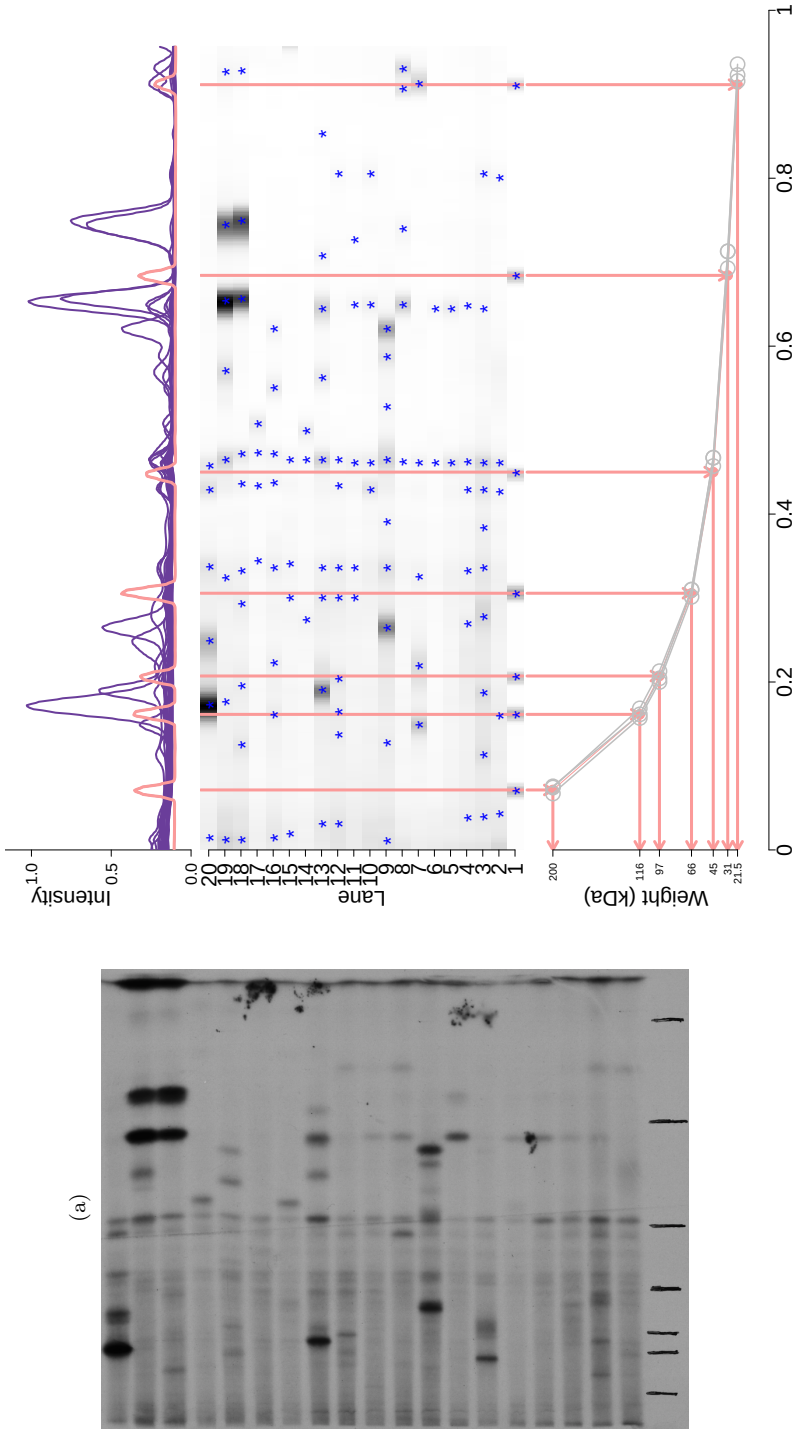

Fig. 2: Gel electrophoresis autoradiography data for 20 samples on one gel.a) Raw GEA image.b) *Top*: Radioactive intensities for all sample lanes;*Middle*: Heatmap of the radioactive intensities by lane (Lane 1 as reference). The blue asterisks (*) denote the detected peaks.*Bottom:* Actual molecular weights (Y-axis) as read from the location along the gel (X-axis). Four location-to-weight curves are shown, each corresponding to reference lane 1s in a gel. Note the marker molecule misalignment.

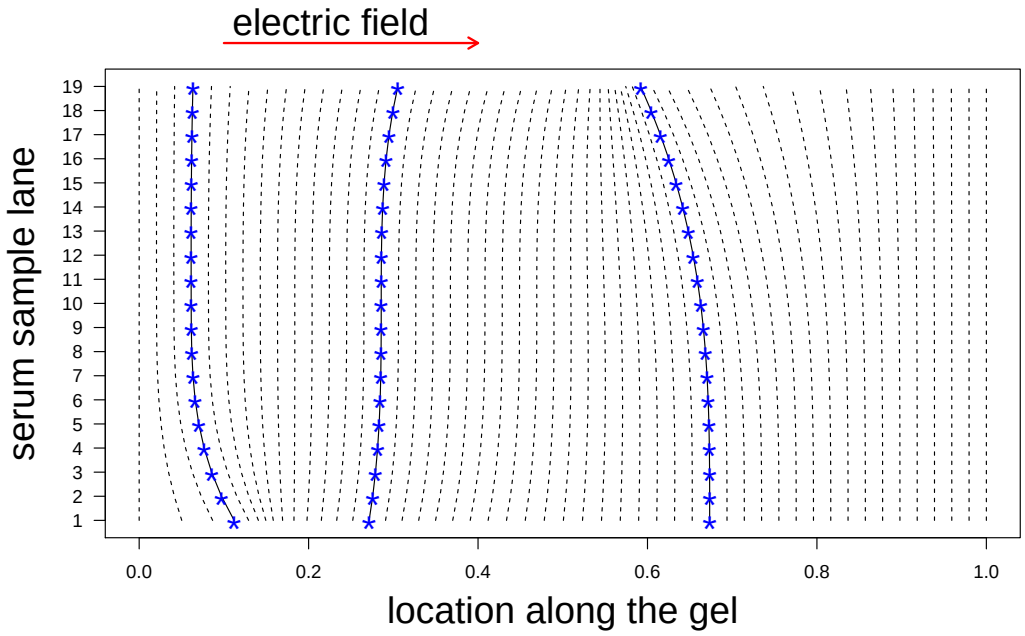

Fig. 3: Example: a gel warping function $\mathcal{S}$ that corrects local L-,S- and 7-shaped stretching or compression. Highlighted are three vertical smooth curves that each aligns the peaks (blue asterisks "$*$") with identical molecular weights.

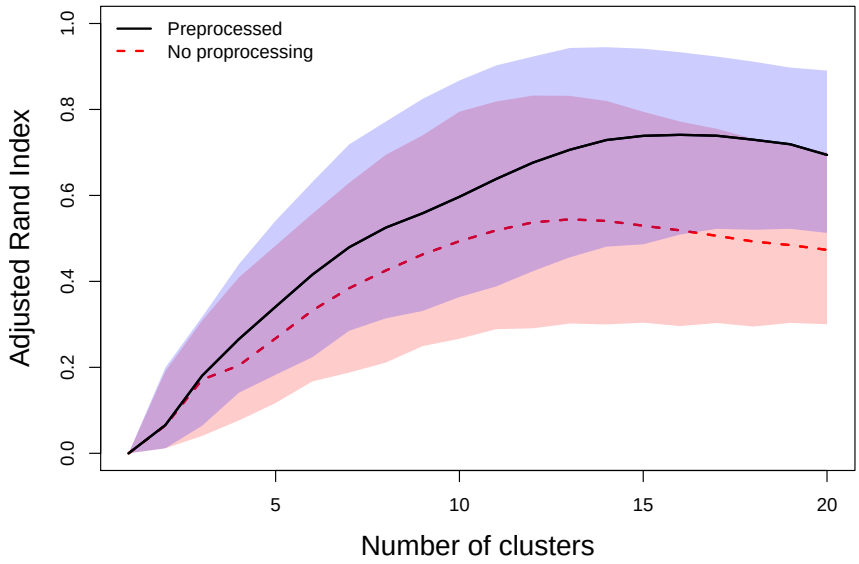

Fig. 4: Adjusted Rand Index $\mathsf{aRI}(\mathcal{C}, \mathsf{True\ Pairs})$ for assessing the similarity between the true clustering and the estimated clustering with ($\mathcal{C} = \hat{\mathcal{C}}(n)$; solid line; blue shaded bands) and without ($\mathcal{C} = \hat{\mathcal{C}}^0(n)$; dashed line; red shaded bands) preprocessing. The central lines and their shaded bands are the mean curves and 95% confidence bands for varying number of clusters n.

Fig. 5: Bayesian gel dewarping result for the second data (reference lane 1s excluded). *Top:* For each gel set, 19 serum lanes at $L = 100$ interior landmarks. Solid blue dots "●" are detected peaks deviating from its true weight. Each detected peak T_{gij} is connected to a red triangle "Δ" that represents the *maximum a posteriori* molecular weight landmark $\hat{Z}_{gij}$. The bundle of black vertical curves visualize the deformations, with each black vertical curve connecting estimated locations with identical molecular weights. The curves are drawn for each landmark. *Bottom:* Marginal posterior probabilities of each landmark protein present in a sample.