Estimating AutoAntibody Signatures to Detect Autoimmune Disease Patient Subsets

Zhenke Wu

Assistant Professor of Biostatistics and

Michigan Institute of Data Science (MIDAS)

University of Michigan

July 10, 2017

zhenkewu.com

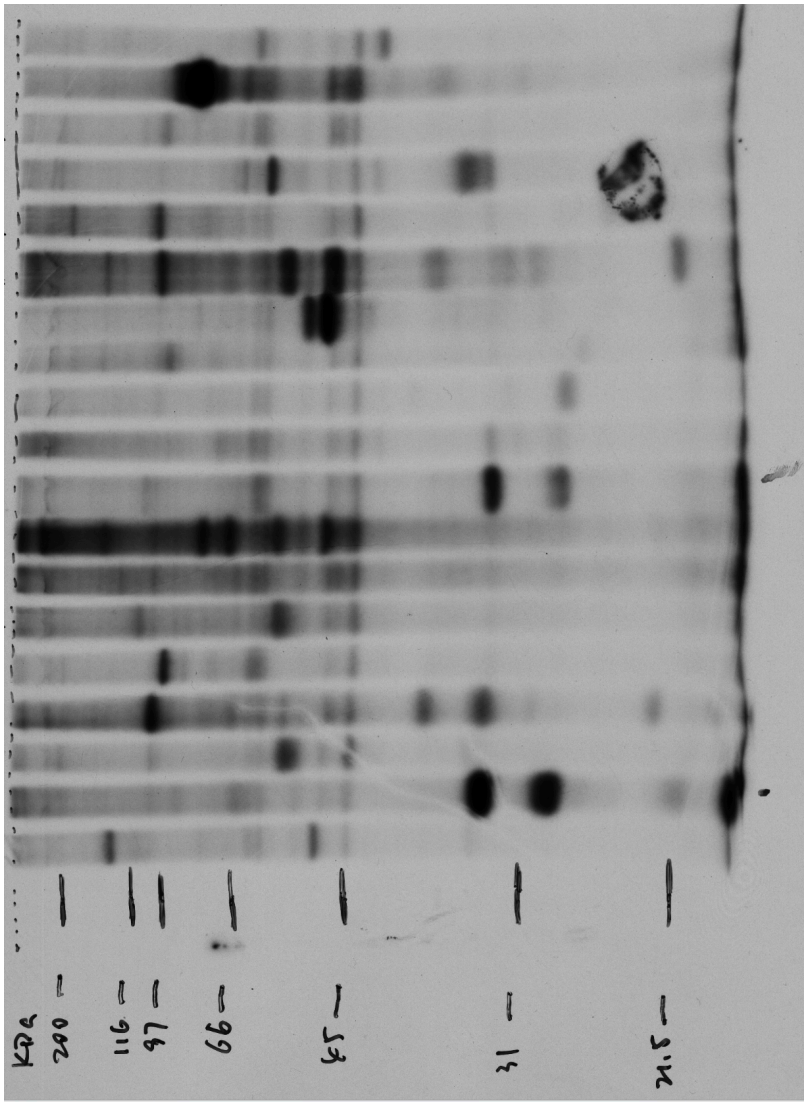
R Package: *spotgear*
https://github.com/zhenkewu/spotgear

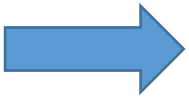
Individualized Health (Precision medicine)

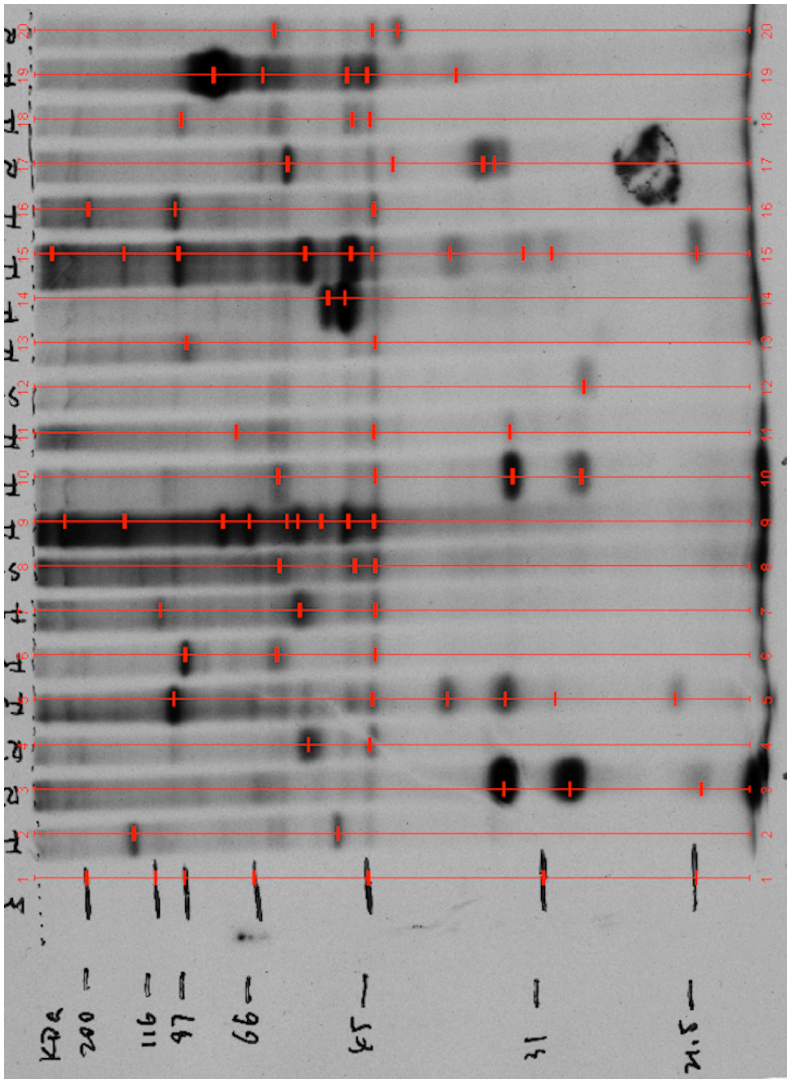
Background

Raw Image:
Gel Electrophoresis Autoradiography; 20 Lanes

Computational Autoradiography and Hand-Picked Bands







Autoimmune Diseases and AutoAntibody Signatures

- Etiology of Autoimmune Diseases: Human immune system's responses to autoantigens; The body produces specific autoantibodies that target these autoantigens but also cause tissue damage

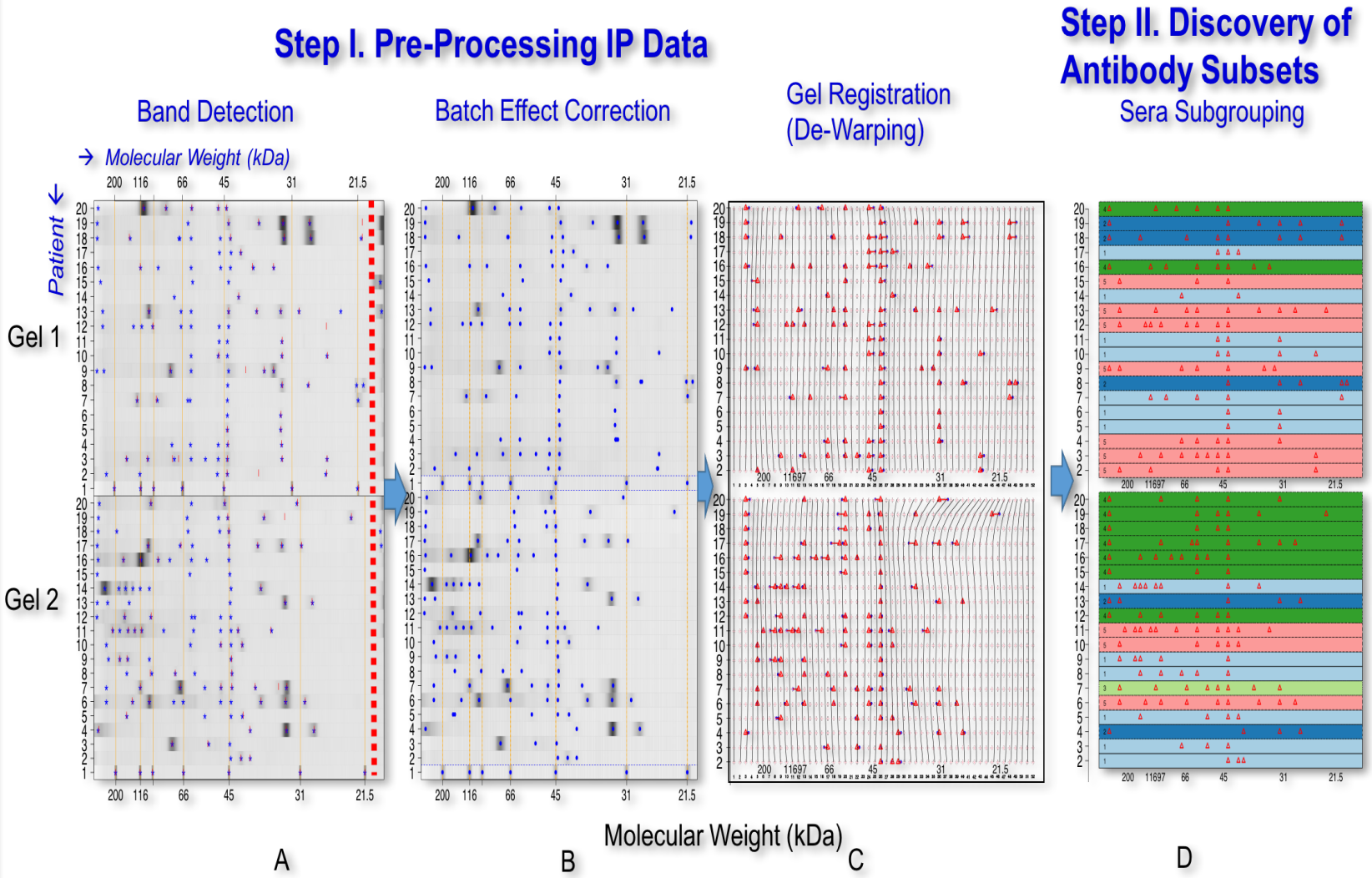
- Heterogeneity: The autoantibody composition is strikingly different among patients

- Measurements: gel electrophoresis autoradiography (GEA)

Gel Electrophoresis Autoradiography (GEA)

- Can generate 100s of possibilities of band patterns
- Gap: Onerous and expensive to validate; Need a method to greatly simplify autoantibody discovery
- Solution: Pre-filtering to define subgroups with similar specificities based on the bands observed by GEA

Individualized Health (Precision medicine)
This Talk: Autoimmune Disease Subsetting
Automated pipeline (Open-Source Software)

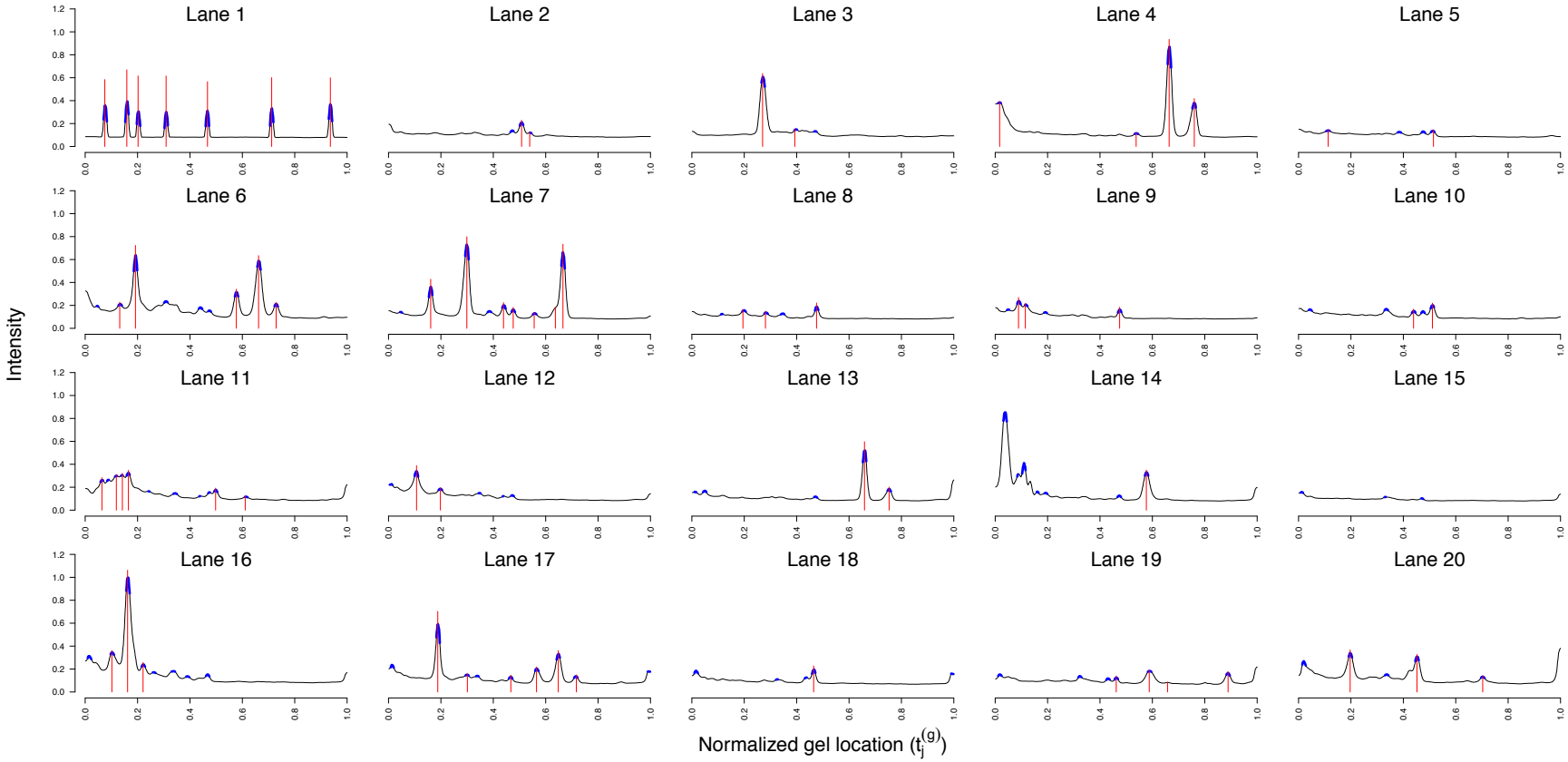


Wu, Casciola-Rosen, Shah, Rosen, Zeger (2017). http://biorxiv.org/content/early/2017/04/21/128199

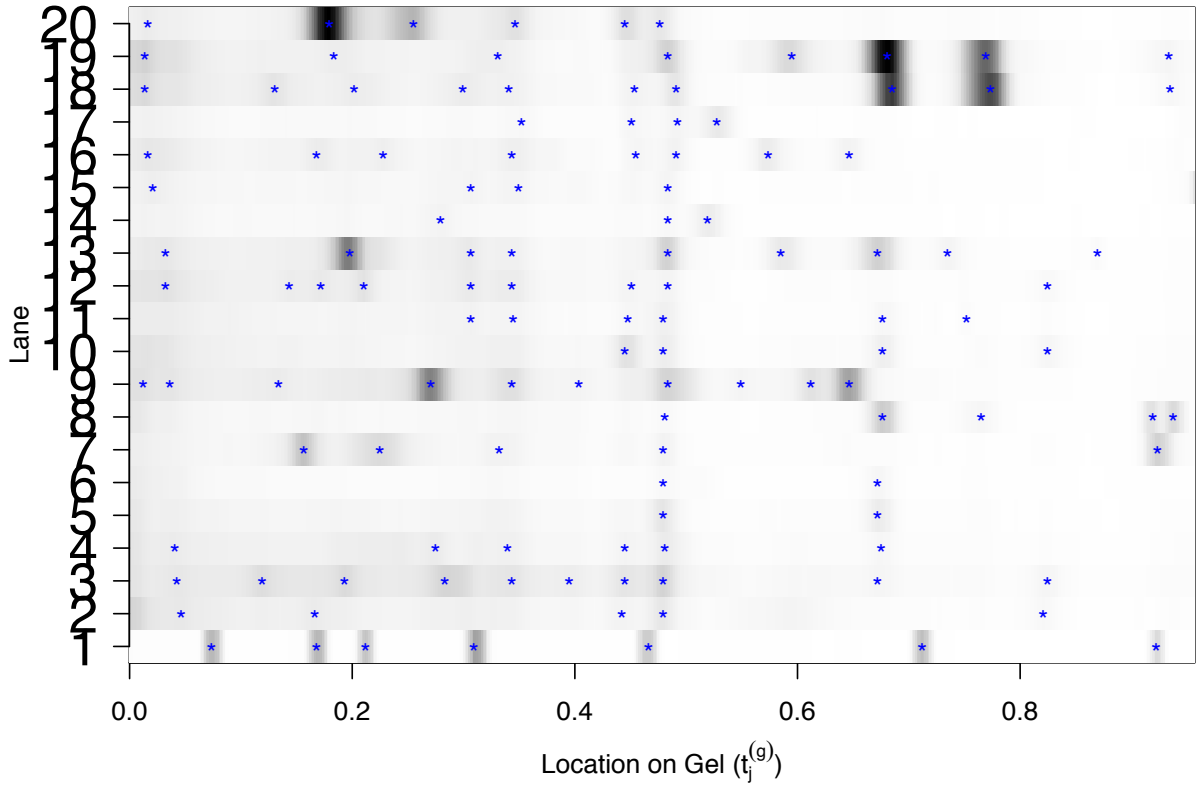
Pre-Processing

Step I-A: Peak Detection

Step I-A: Peak Detection



Step I-A: Automated Peak Detection (Overlaid against gel image; "*" detected peaks)



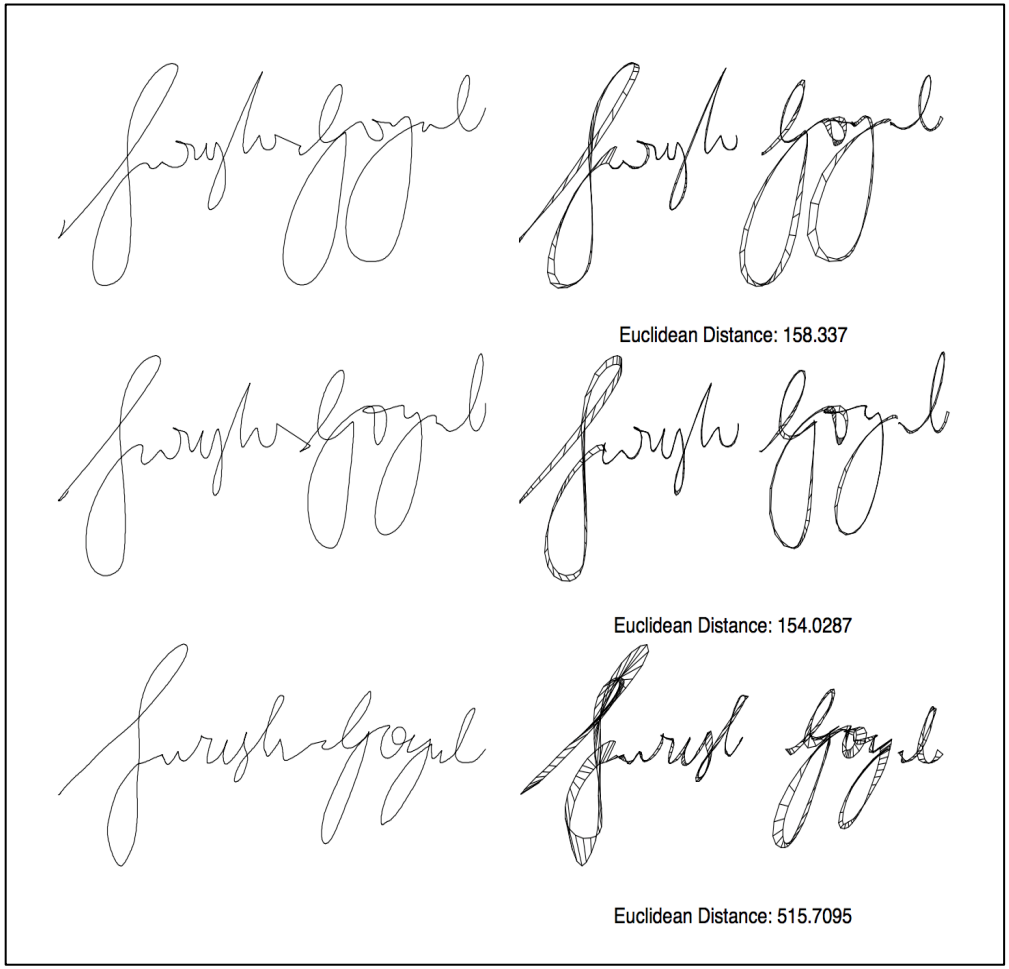
u_{gi}: Lane number for Lane $i = 1, \ldots, N_g$, Gel $g = 1, \ldots, G$

T_{gij}: Location for the jth peak (counted from the left; j-th "*", $j = 1, \ldots, J_{gi}$, for lane i, gel g)

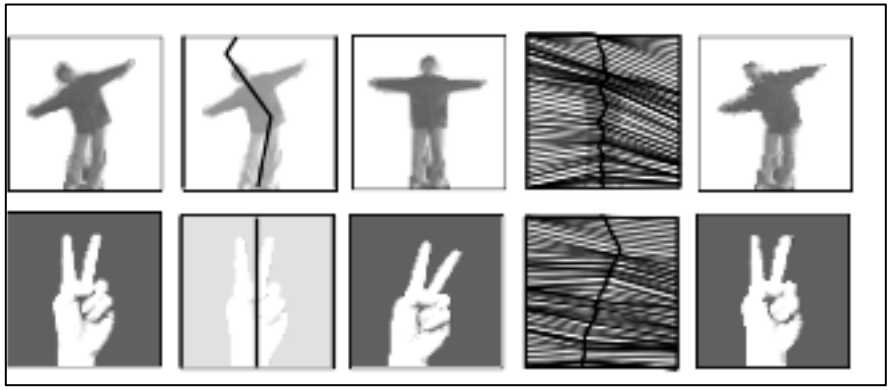
Step I-B: Batch Effect Correction

Piecewise Linear Warping by Reference Lanes

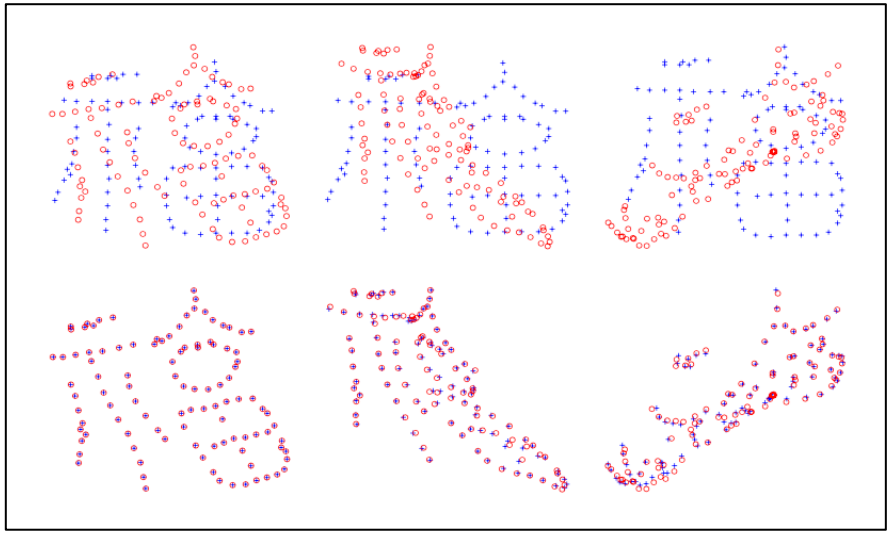
Warping: Examples



Signature Verification: Hastie et al. 1991

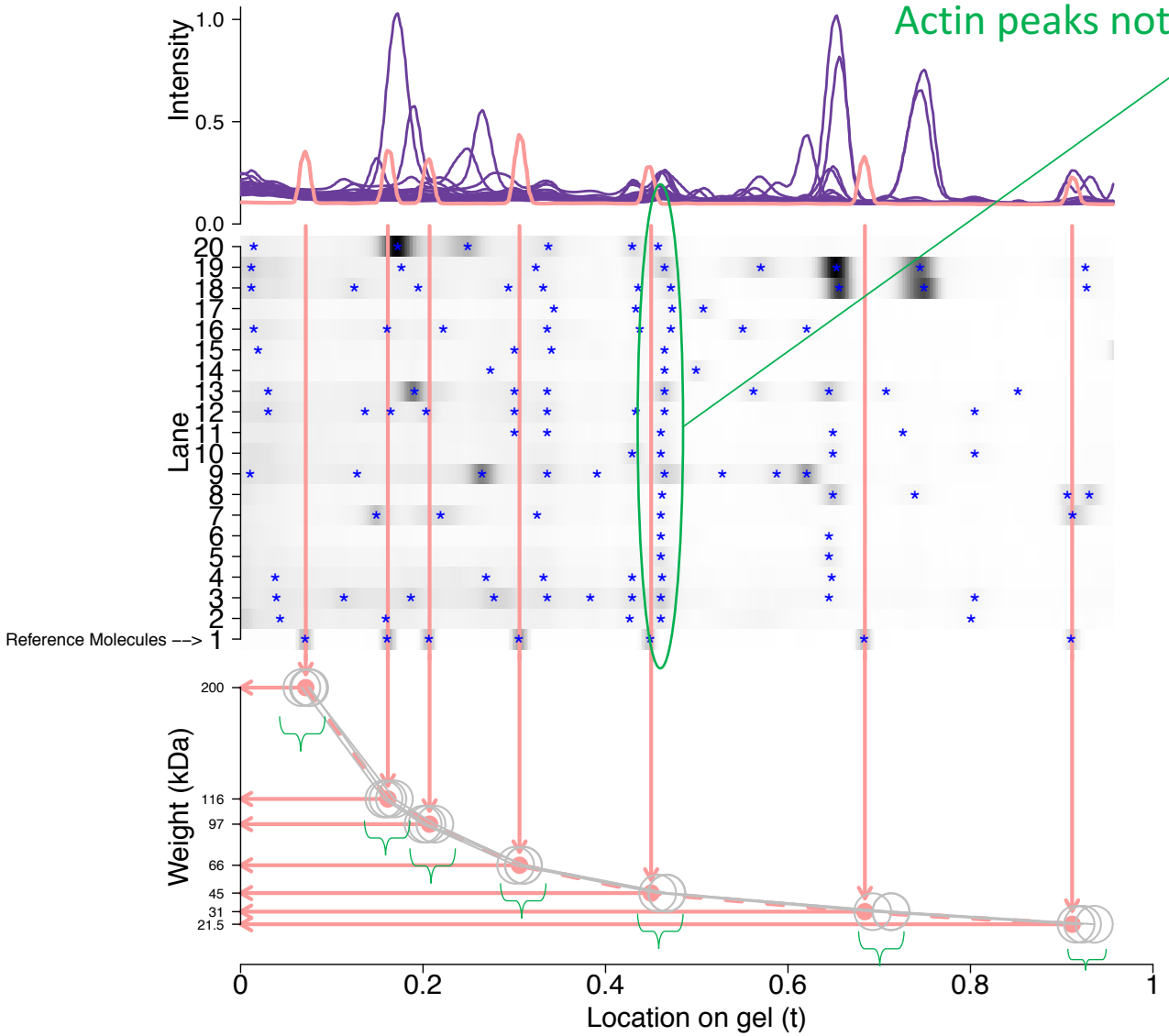


Motion Alignment: Uchida and Sakoe, 2000



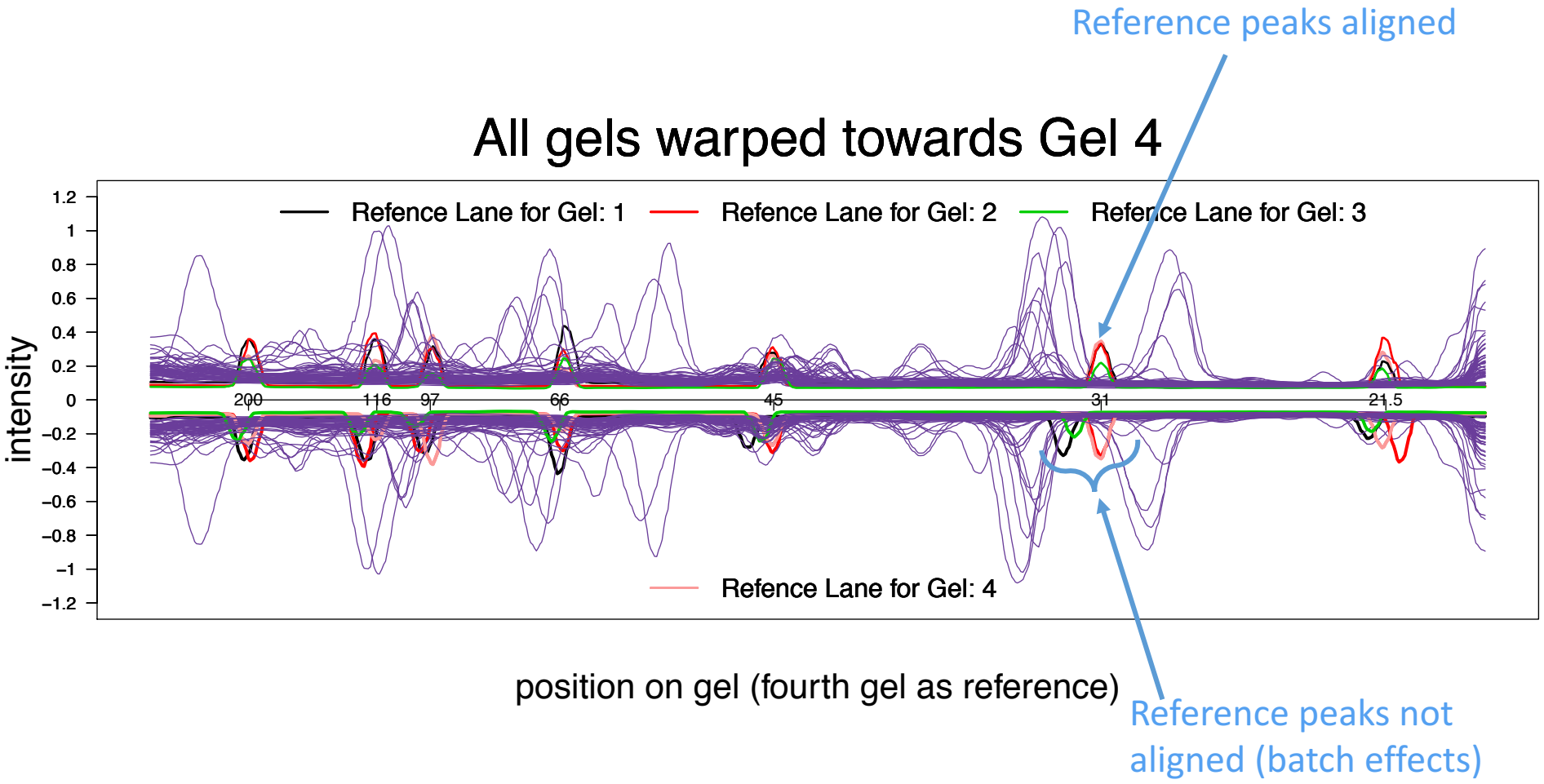
Handwritten Chinese: Ma and Zhao (2015)

Batch effect and Warping
Must address before meaningful subgrouping



Reference proteins not aligned: Batch effects!

Step I-B: Batch Effect Correction: Piecewise Linear Warping by Reference Lanes



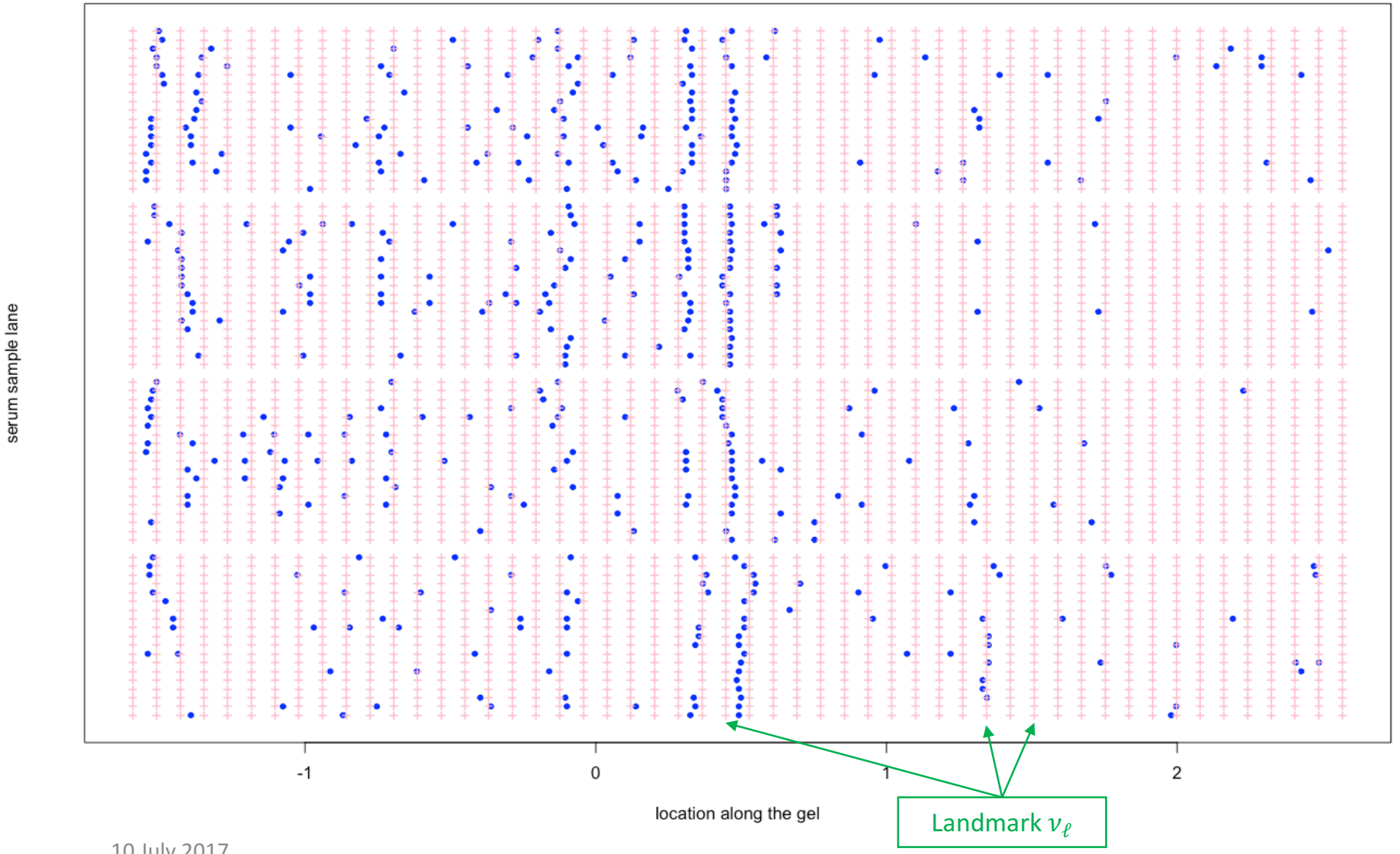
Step I-C: Image Registration:

Two-Dimensional De-Warping

Step I-C: Two-Dimensional De-Warping

- The physical process of autoradiography could cause image deformation
- Challenges
  - In general, few light-weight proteins on the right side of the image; If we don't see bands, how to align? *Solution*: align to a grid of protein landmarks and stretch/compress
  - Ubiquitous proteins on multiple gels must be aligned (e.g., actin) *Solution*: Discretized non-homogeneous Poisson process with shared intensity across gels
  - The observed peak locations are noisy. *Solution*: Gaussian mixture model with a mean determining the true location.

Step I-C: Align the peaks – "where do the peaks belong?": (50 protein landmarks; Pink "+")



Step I-C: 2-Dimensional Image Registration: Down-sampling to Landmarks (50 for each lane; Pink +)

- Peak-to-landmark Indicators:
  1. $Z_{gij}, \in \{1, \ldots, L\}, j = 1, \ldots, J_{gi}$ (match a blue "*" to a pink "+") e.g, $Z_{gij} = 3$ means the peak is matched to Landmark 3
  2. Constrain $Z_{gi,j-1} \leq Z_{gij}$ to prevent reverse matching

- Bayesian Model for Aligning Peaks to Landmarks

Cumulative intensity; Controls the average total number of peaks

$$\text{Number of observed peaks in lane } i, \text{ gel } g: \quad J_{gi} \overset{d}{\sim} \text{Poisson}(\Lambda),$$

$$\text{Peak} - \text{to} - \text{landmark indicators}:$$

$$(Z_{gi1}, \ldots, Z_{giJ_{gi}}) = \text{ increasing sort } \{Z^*_{gi1}, \ldots, Z^*_{giJ_{gi}}\},$$

$$Z^*_{gij} \overset{iid}{\sim} \text{Categorical}\left(\{\lambda^*_\ell\}_{\ell=1}^{L}\right),$$

Landmark-specific intensity; Independent of g; Hence, when possible, encourages nearby peaks to be aligned to an **identical** landmark

Step I-C: 2-Dimensional Image Registration: Gaussian Mixture Model for Noisy Peak Locations

- Given the peak-to-landmark indicators Z_{gij}, model the observed peaks T_{gij} as observations from a Gaussian mixture model with *L* components:

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

- $\mathcal{S}_g : \mathbb{R}^2 \to \mathbb{R}$, warps landmarks: $(\nu_\ell, u_{gi}) \mapsto (\mathcal{S}_g(\nu_\ell, u_i), u_{gi})$.
- The set $\mathcal{I}_{gij}(\nu_\ell, A_0) \triangleq \{t : |t - \nu_\ell| < A_0 \text{ and } t > T_{gi,j-1}\}$ assumes a peak appears within distance A_0 from its true landmark
- Let P_g the peaks for gel *g;* let P *collect all the peaks*

Step I-C: 2-Dimensional Image Registration: Warping Function by Tensor Product Basis Expansion

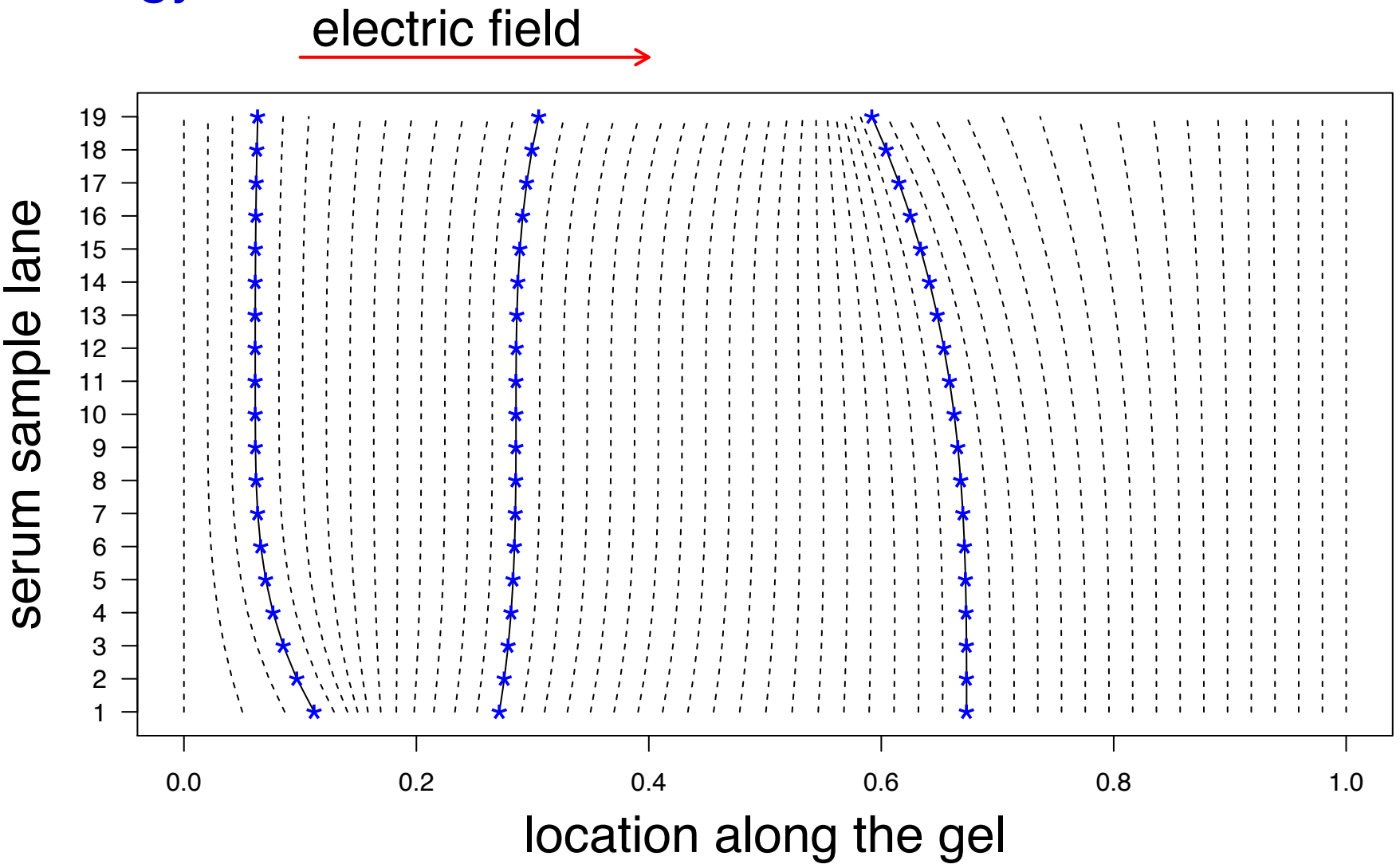
$$\mathcal{S}_g(\nu, u) = \sum_{s=1}^{T_\nu} \sum_{t=1}^{T_u} \beta_{gst} B_{g1s}(\nu) B_{g2t}(u)$$

- $B_{1s}(\cdot)$ and $B_{1t}(\cdot)$: s-th and t-th cubic B-spline basis along two coordinate directions, respectively

- Implementing Warping function Constraints and Priors
  - Boundary constraint: $\mathcal{S}_g(\nu_0, u) = \nu_0, \mathcal{S}_g(\nu_{L+1}, u) = \nu_{L+1}$
  - Smoothness: Bayesian Penalized-Splines to make adjacent $\{\beta_{gst}\}$ similar
  - Monotonic warping: $\nu_0 < \mathcal{S}_g(\nu_{\ell-1}, u) < \mathcal{S}_g(\nu_\ell, u) < \nu_{L+1}, \forall \ell = 1, \ldots, L, \forall u$
  - Vary by gel: $\mathcal{S}_g(u, \nu_\ell)$
  - Can be implemented via constraints on $\{\beta_{gst}\}$

Key reference: Stefan and Brezger, 2004

Step I-C: A Mathematical Model for Warping
Strategy: Estimate the model, then reverse



Step I-C: Goal of 2-Dimensional Image De-warping

- The posterior distribution $[\mathbf{Z} \mid P]$

Step I-C: Posterior Inference of the De-Warping

$$\prod_{g=1}^{G}\Bigg\{\underbrace{\prod_{i=1}^{N_g}\Bigg[\prod_{j=1}^{J_{gi}} N\left(T_{gij};\boldsymbol{B}_{g1}(\nu_{Z_{gij}})'\boldsymbol{\beta}_g\boldsymbol{B}_{g2}(u_{gi}),\sigma_\epsilon^{-2}\right)\mathbf{1}\{T_{gij}\in\mathcal{I}_{gij}(\nu_{Z_{gij}},A_0)\}}_{\text{likelihood (2.2)}}$$

$$\times\underbrace{J_{gi}!\prod_{j=1}^{J_{gi}}\text{Categorical}(Z_{gij};\boldsymbol{\lambda})\mathbf{1}\{Z_{gij}\le Z_{gi,j+1},j=1,\ldots,J_{gi}-1\}\Bigg]}_{\text{prior of }\mathbf{Z}}$$

$$\times\underbrace{N_{T_\nu-1}\left(\{\beta_{gs1}\}_{s=1}^{T_\nu-1};\boldsymbol{\beta}^{\text{id}}_{[-T_\nu]},\sigma_{g1}^{-2}\Delta_1'\Delta_1\right)\mathbf{1}\{\nu_0=\beta_{g11}<\ldots<\beta_{gs1}<\ldots<\beta_{g,T_\nu-1,1}<\nu_{L+1}\}\cdot p(\sigma_{g1}^2)}_{\text{prior (2.6) and hyperprior of the smoothing parameter}}$$

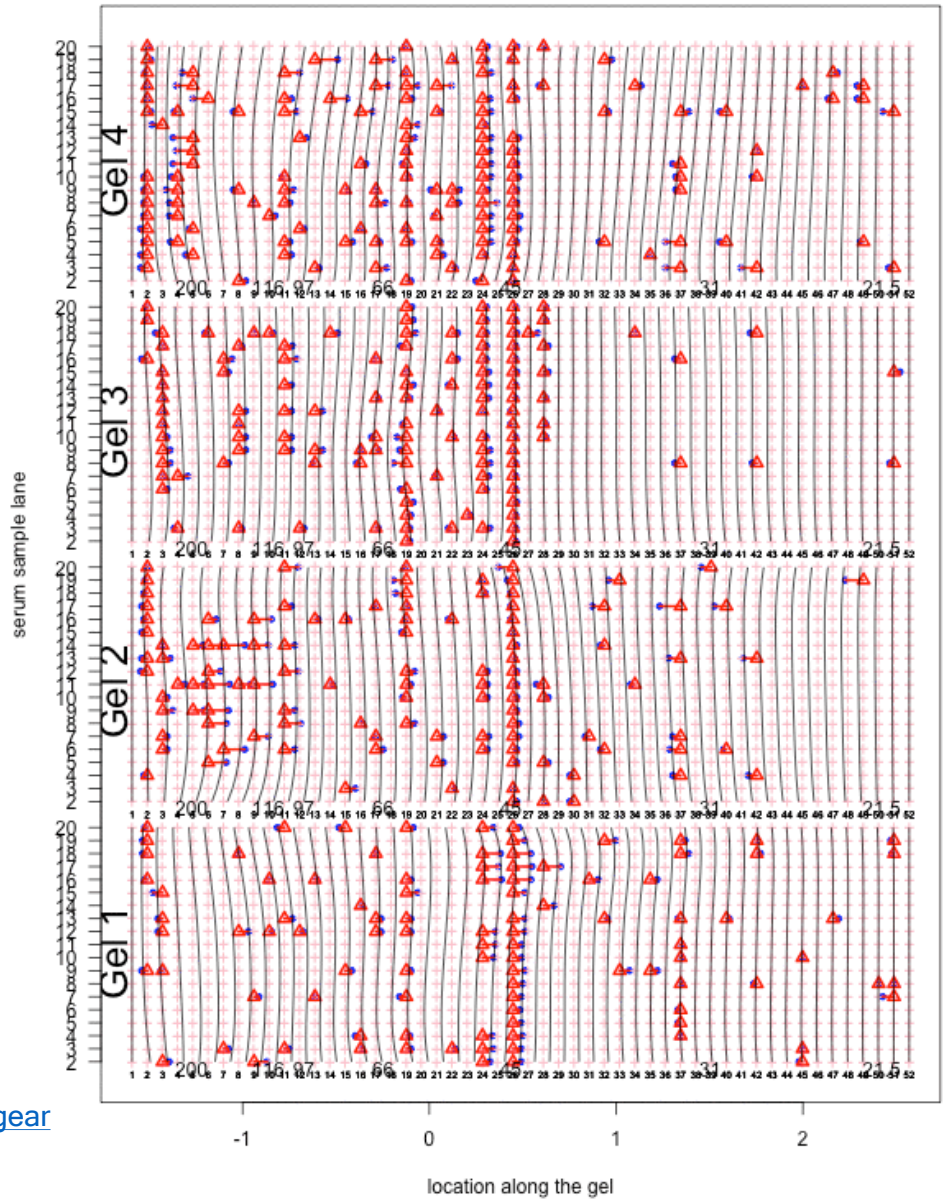
$$\times\underbrace{\prod_{s=2}^{T_\nu-1}\left[N_{T_u}\left(\{\beta_{gst}\}_{t=1}^{T_u};\mathbf{0},\sigma_{gs}^{-2}\Delta_2'\Delta_2\right)\cdot p(\sigma_{gs}^2,\rho_g)\right]\Bigg\}}_{\text{prior (2.7) and hyperpriors of the smoothing parameters}}\times\underbrace{p(\boldsymbol{\lambda})}_{\text{hyperprior for }\mathbf{Z}},\qquad(2.9)$$

1. Joint distribution (data + unknowns) → Posterior distribution of unknown given data: e.g., $[\mathbf{Z} \mid P]$, the posterior of the peak alignment indicators
2. Tool: Markov chain Monte Carlo (MCMC)
3. Idea: Simulate samples from the joint posterior distribution of the unknowns

Reference: Wu, Casciola-Rosen, Shah, Rosen, Zeger (2017).
http://biorxiv.org/content/early/2017/04/21/128199

Step I-C: Align the peaks -- Result

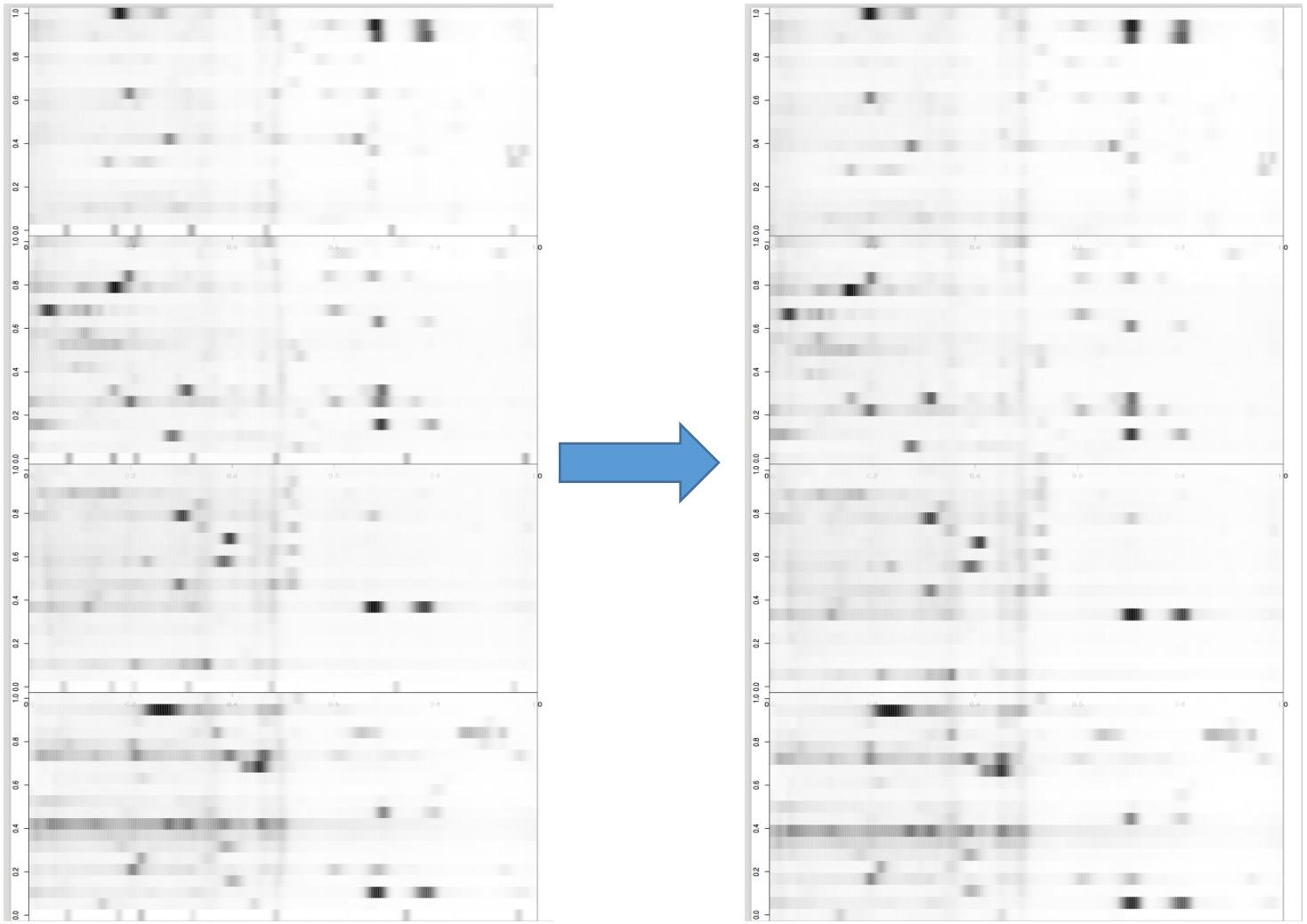
Animation; Δ for signature; "●" for observed peaks



https://github.com/zhenkewu/spotgear

10 July 2017

Step I-C: Alignment
Before and After

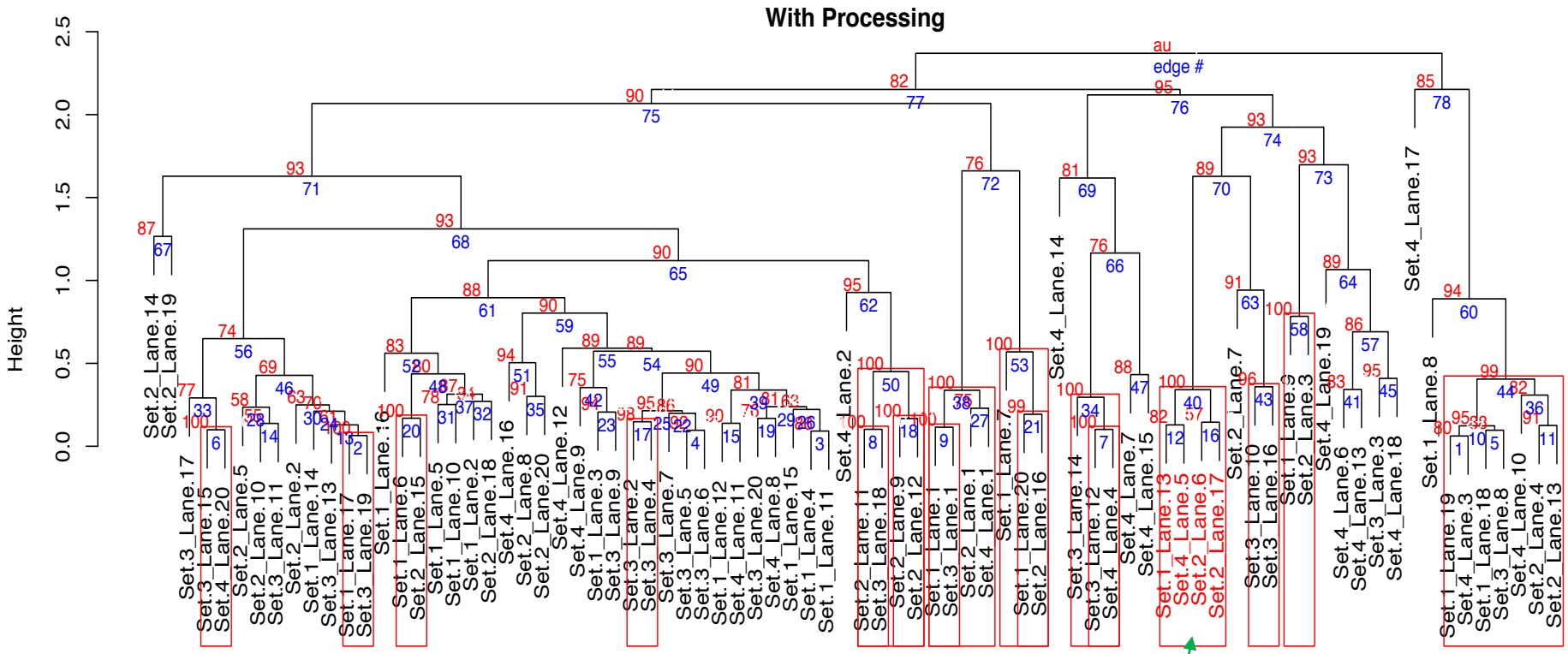


Step II: Cluster the Pre-processed High-Frequency Data

Hierarchical Clustering

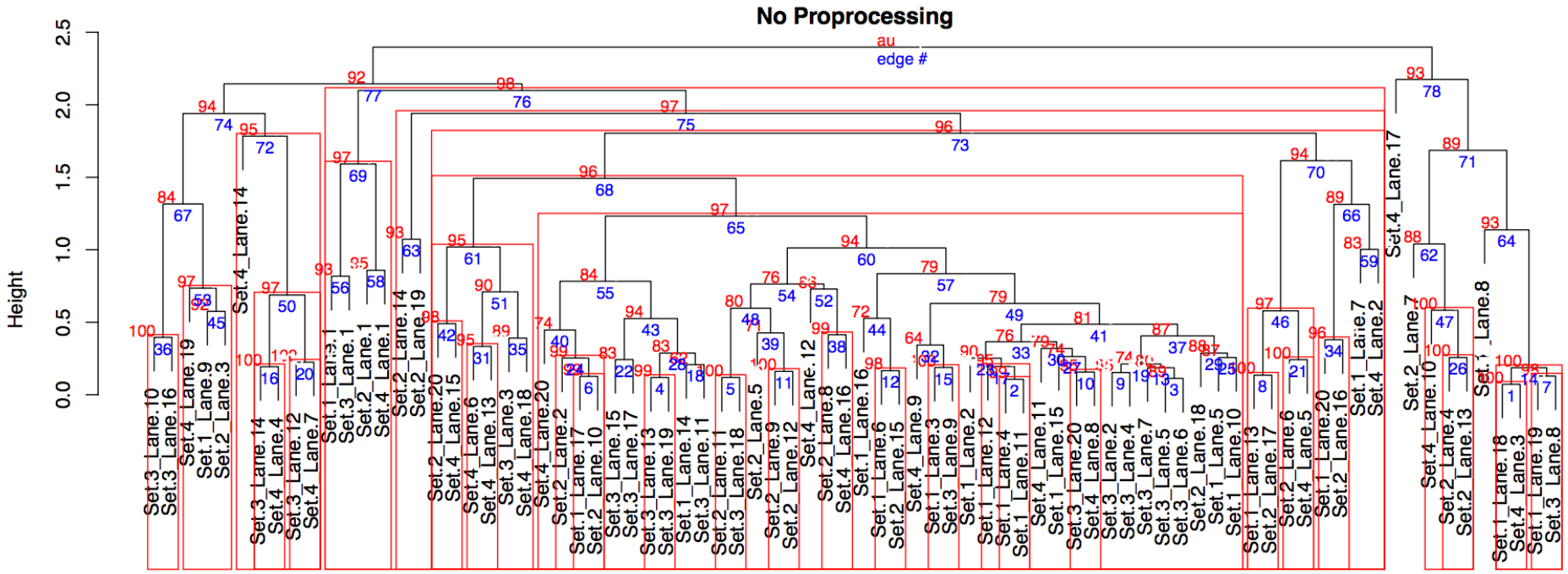
Step II: Subgrouping Results from 80 Pre-processed High-Frequency Data

Adjacent in the tree → similar in autoantibody signatures



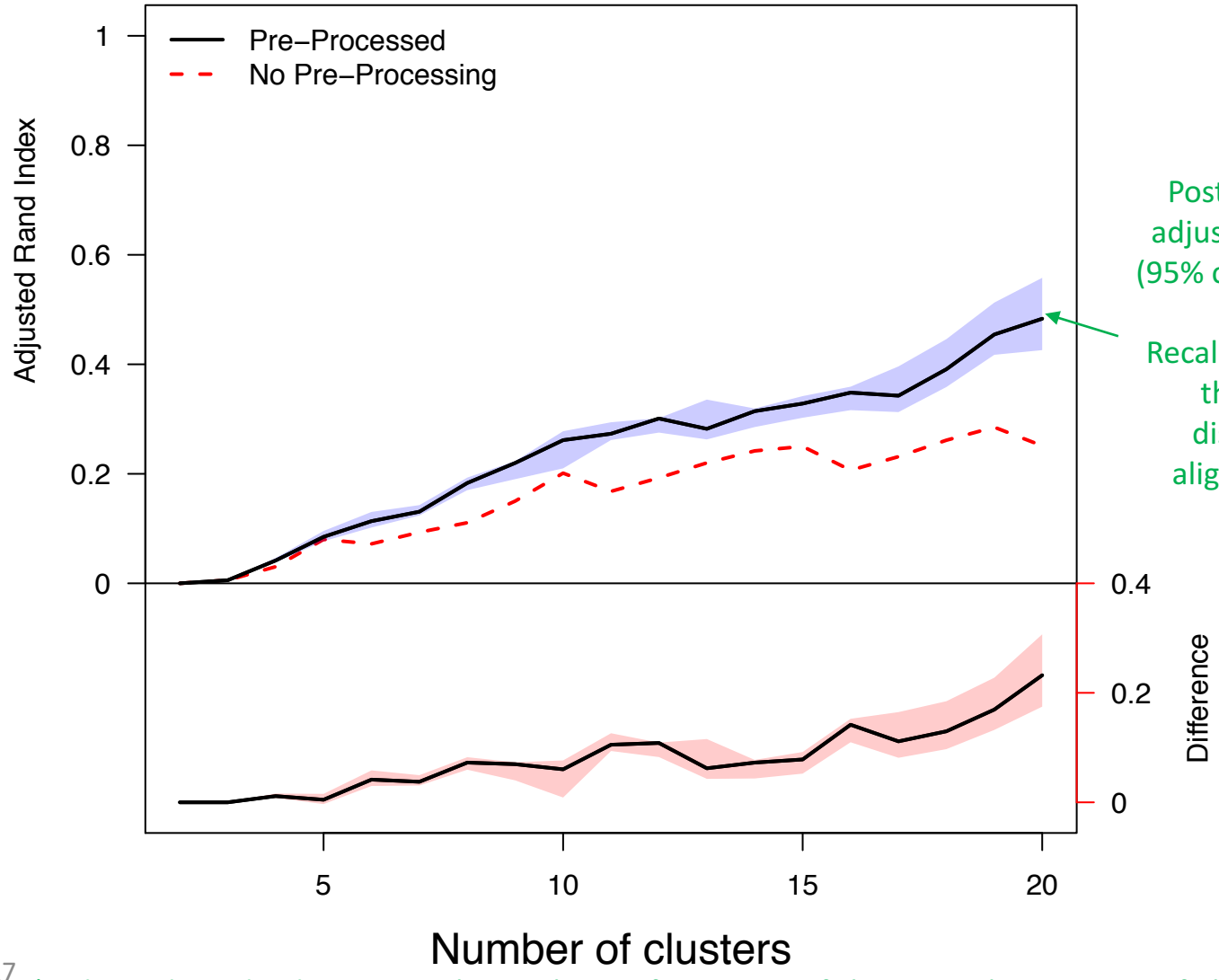
Validated cluster:
contain antibodies against “anti-PMScL”

Step II: Subgrouping Results from 80 *Raw* High-Frequency Data



What's the difference compared to the one WITH pre-processing? Why?

Step II: Comparison between the adjusted Rand indices obtain with and without pre-processing (A separate experiment with replicates; 20 sample, long- and short- exposures)



Posterior mean of adjusted Rand index (95% credible interval)

Recall that we obtain the posterior distribution of alignment $[\mathbf{Z} \mid P]$

* Adjusted Rand index: assess the similarity of two ways of clustering the same set of observations

Summary

- Problem: Human recognition of autoantibody patterns and hence clustering becomes more difficult when patterns are composite and on multiple gels

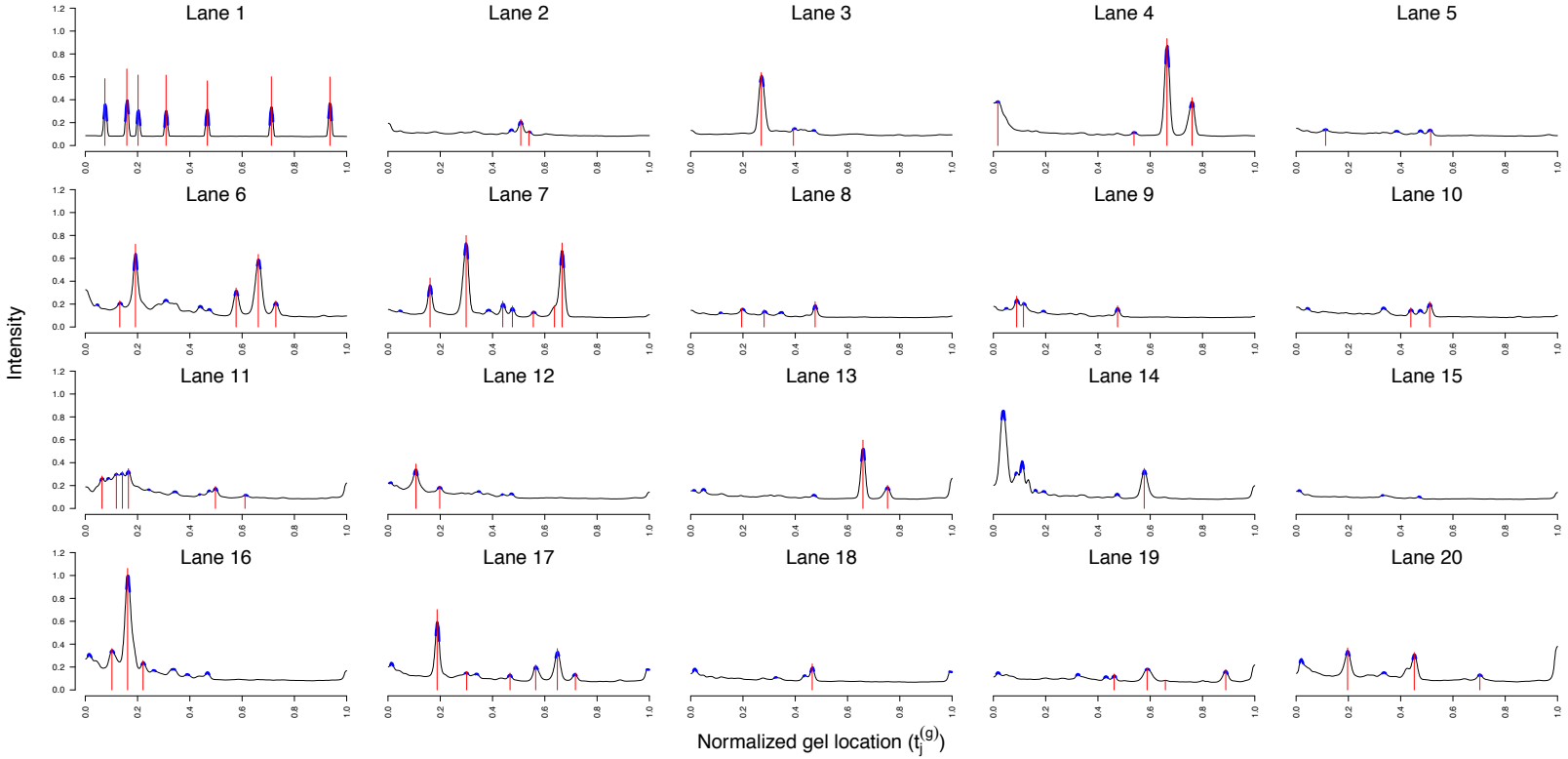
- Method: novel automated algorithms that
  - Estimate autoantibody signatures
  - Improve accuracy of subgroup discovery

- Free Public Open-Source Software: https://github.com/zhenkewu/spotgear

- Manuscript: Wu, Casciola-Rosen, Shah, Rosen, Zeger (2017). http://biorxiv.org/content/early/2017/04/21/128199

- Ongoing work: novel Bayesian clustering model to find disease subsets informed by the biology that autoantibodies in the cells work in complexes

Step I-A: Peak Detection



1. Local Difference Scoring:

$$\text{score}_{gi}(b) = \text{sign}\left\{M^0_{gib} - M^0_{gi,\ell(b)}\right\} + \text{sign}\left\{M^0_{gib} - M^0_{gi,r(b)}\right\} + \text{sign}\left\{M^0_{gib} - \min_{\ell(b)\le b'\le r(b)} M^0_{gib'} - C_0\right\},$$

A left bin: 10 bins away

A right bin: 10 bins away

Minimum elevation

2. Peak calling: find contiguous high-scoring regions and then find their respective centers