Estimating AutoAntibody Signatures to Detect Autoimmune Disease Patient Subsets

Zhenke Wu

Assistant Professor
Department of Biostatistics, University of Michigan

27 July 2017

The 62nd Annual International Biometric Society Meeting of the Brazilian Region (RBras 2017)

R package: spotgear

https://github.com/zhenkewu/spotgear

Common Questions on Individual and Population Health

1. a. What is the person's health state given health measurements?
   b. What is the population distribution of health states?
   (Wu et al., 2015, *JRSS-C*; Wu and Zeger, 2016a,b)
2. a. What is the person's health trajectory?
   b. What is the population's characteristics of health trajectory?
3. Does a particular intervention improve health - on average/for a particular person? (Wu et al., 2014, *Biometrics*; Frangakis, Qian, Wu, Diaz, 2015, *Biometrics*)
4. Are interventions being used optimally?

Example I

Pneumonia Etiology Research for Child Health (PERCH)

Background:

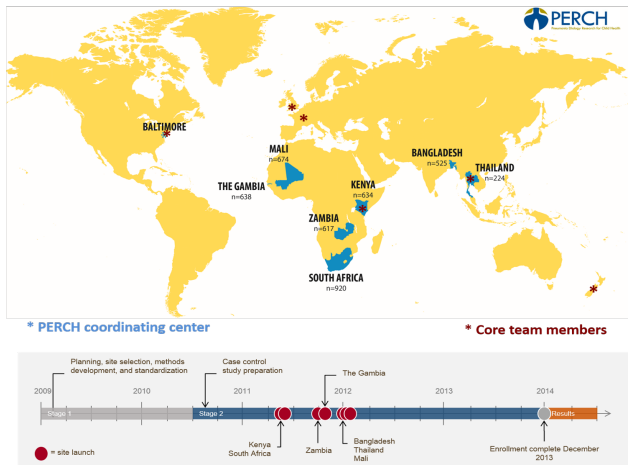
- $>$ 30 possible infectious causes
- Difficult to directly observe

Goal:

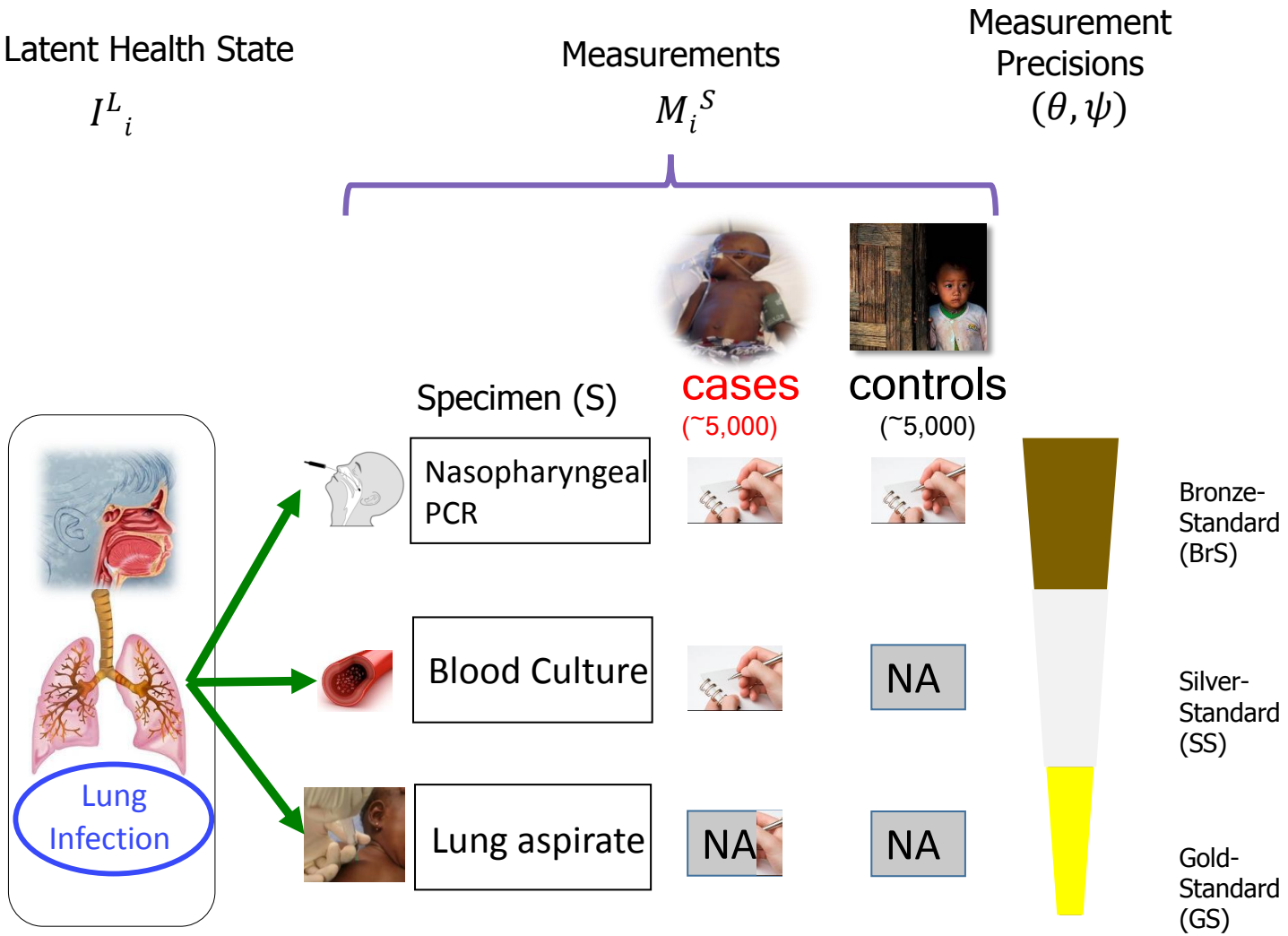
- Population disease etiology estimation
- Individual diagnosis

Study details:

- \$40-mil, Gates-funded 7-country study; Sites at Sub-Saharan Africa and South Asia
- Diverse measures; variable precisions
- $\sim$5,000 cases and $\sim$5,000 controls



Measurements of Different Quality

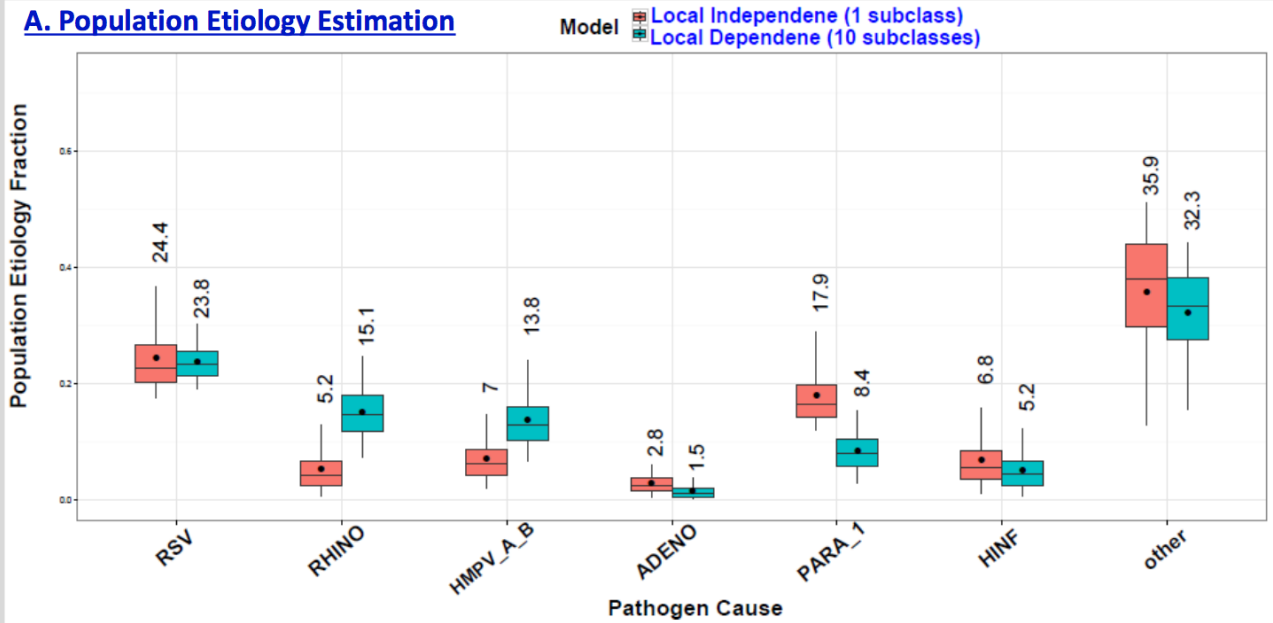


---

*NP: nasopharyngeal; PCR: polymerase chain reaction; LA: lung aspirate

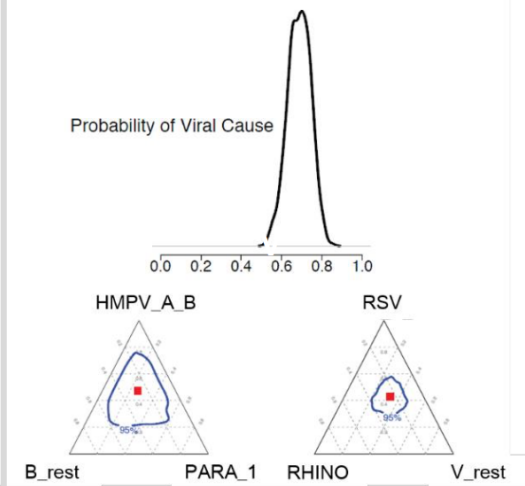
Nested Partially-Latent Class Models for Population and Individual Estimations

A. Population Etiology Estimation



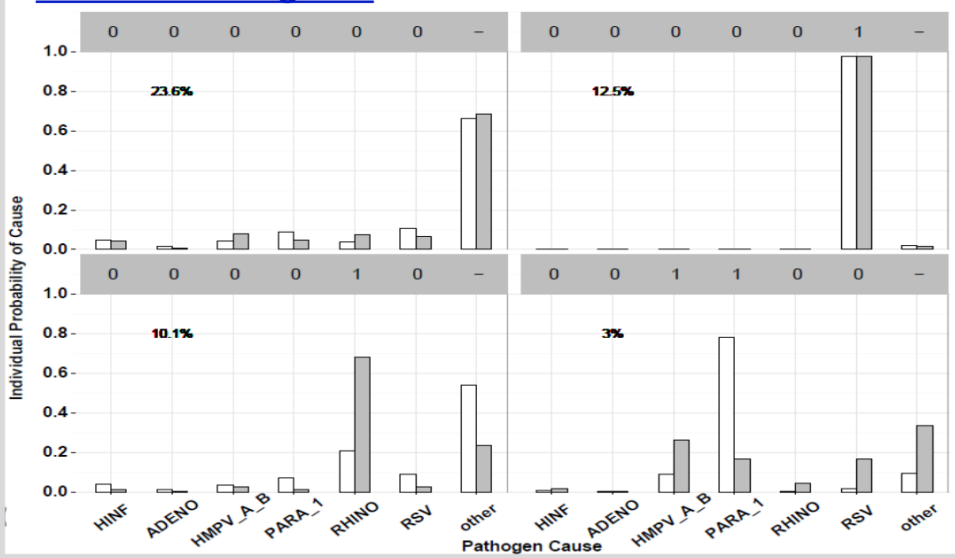
Marginal posterior distributions of the population etiology for 6 leading pathogens plus other. Results based on two models are shown.

B. Estimation by Pathogen Taxonomy



Posterior of viral and bacterial etiology

C. Individual Diagnosis

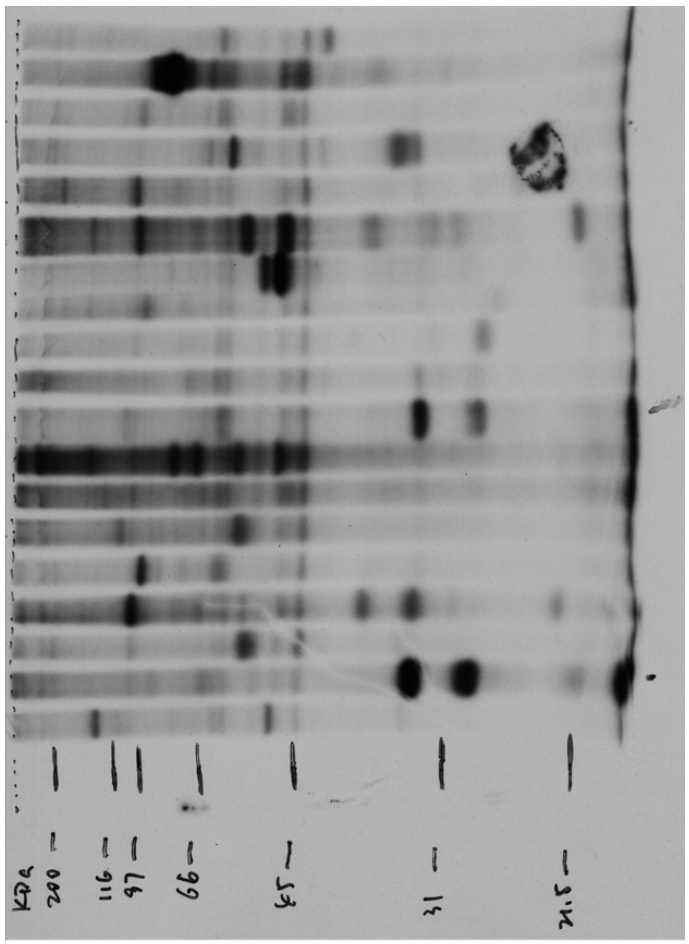


Individual predictions of the cause of pneumonia for 4 patterns of NPPCR data (binary codes in gray bands with its observed frequency shown below). Results from two models are shown. The top two have similar predictions; The bottom two differ, depending on how we correct for false positives.

Example II: Raw Data

Gel Electrophoresis Autoradiography; 20 Samples

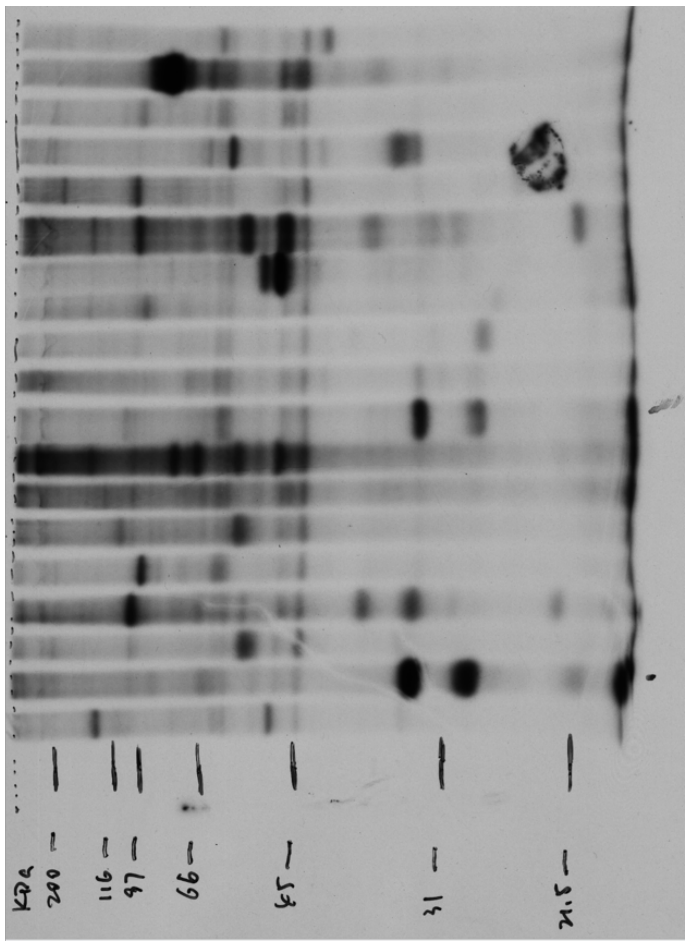
Raw Image



Example II: Raw Data

Gel Electrophoresis Autoradiography; 20 Samples

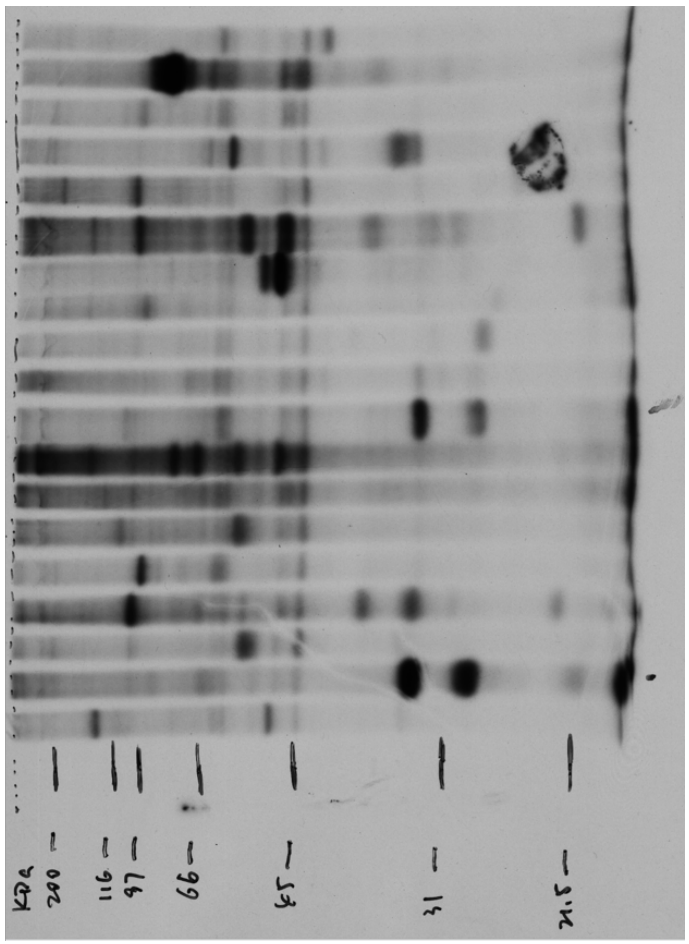
Raw Image



Example II: Raw Data

Gel Electrophoresis Autoradiography; 20 Samples

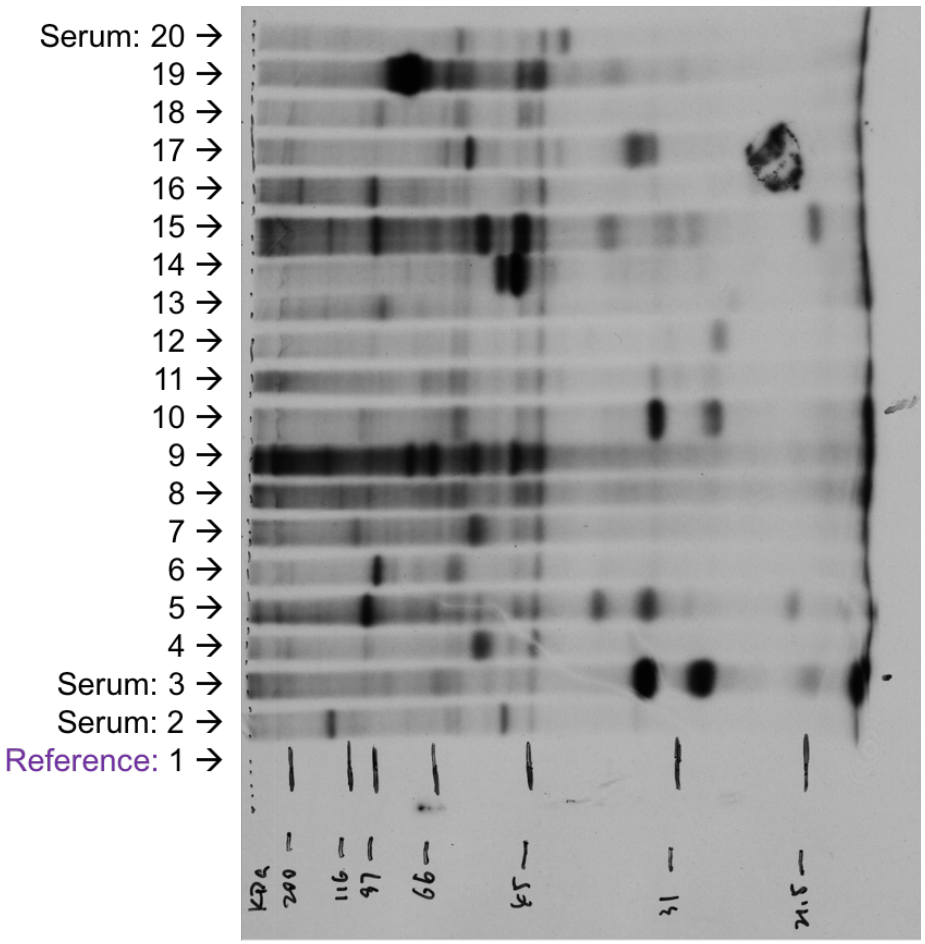
Raw Image



Example II: Raw Data

Gel Electrophoresis Autoradiography; 20 Samples

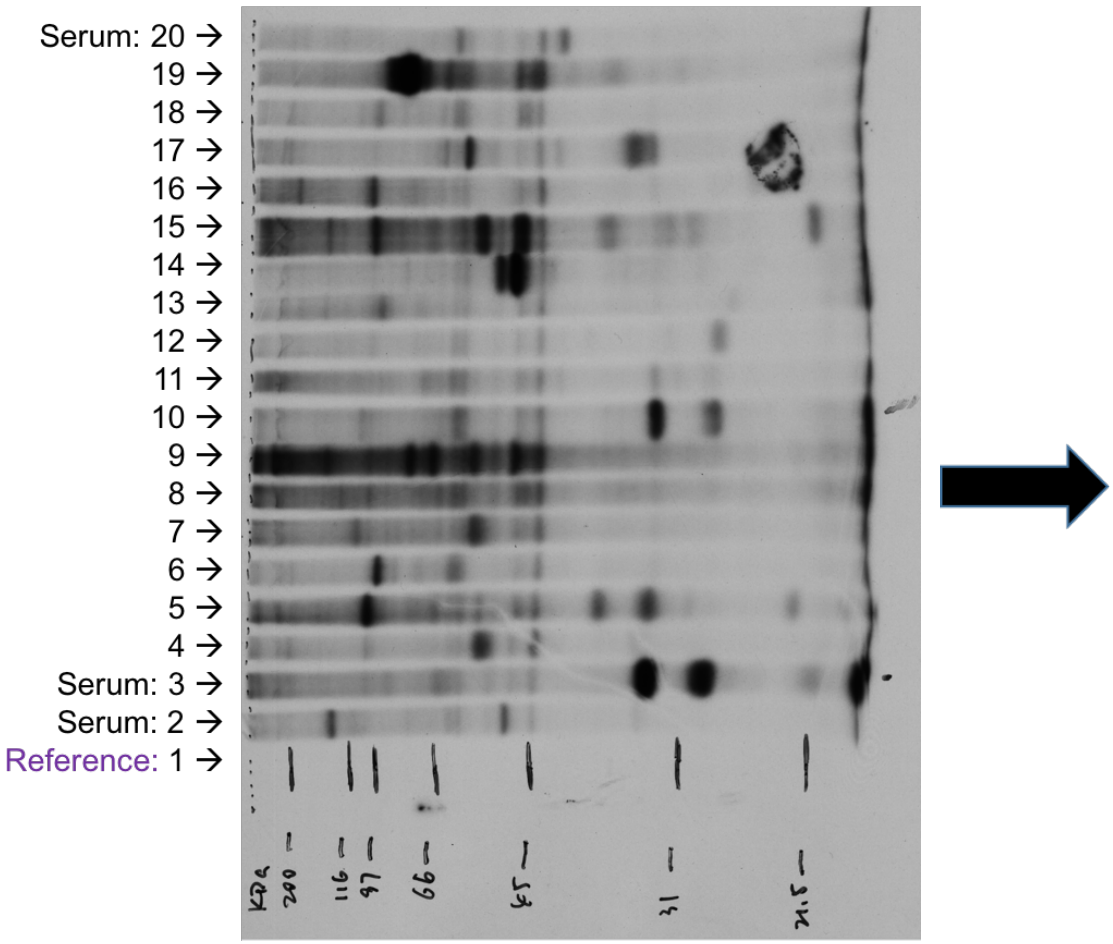
Raw Image



Example II: Raw Data

Gel Electrophoresis Autoradiography; 20 Samples

Raw Image

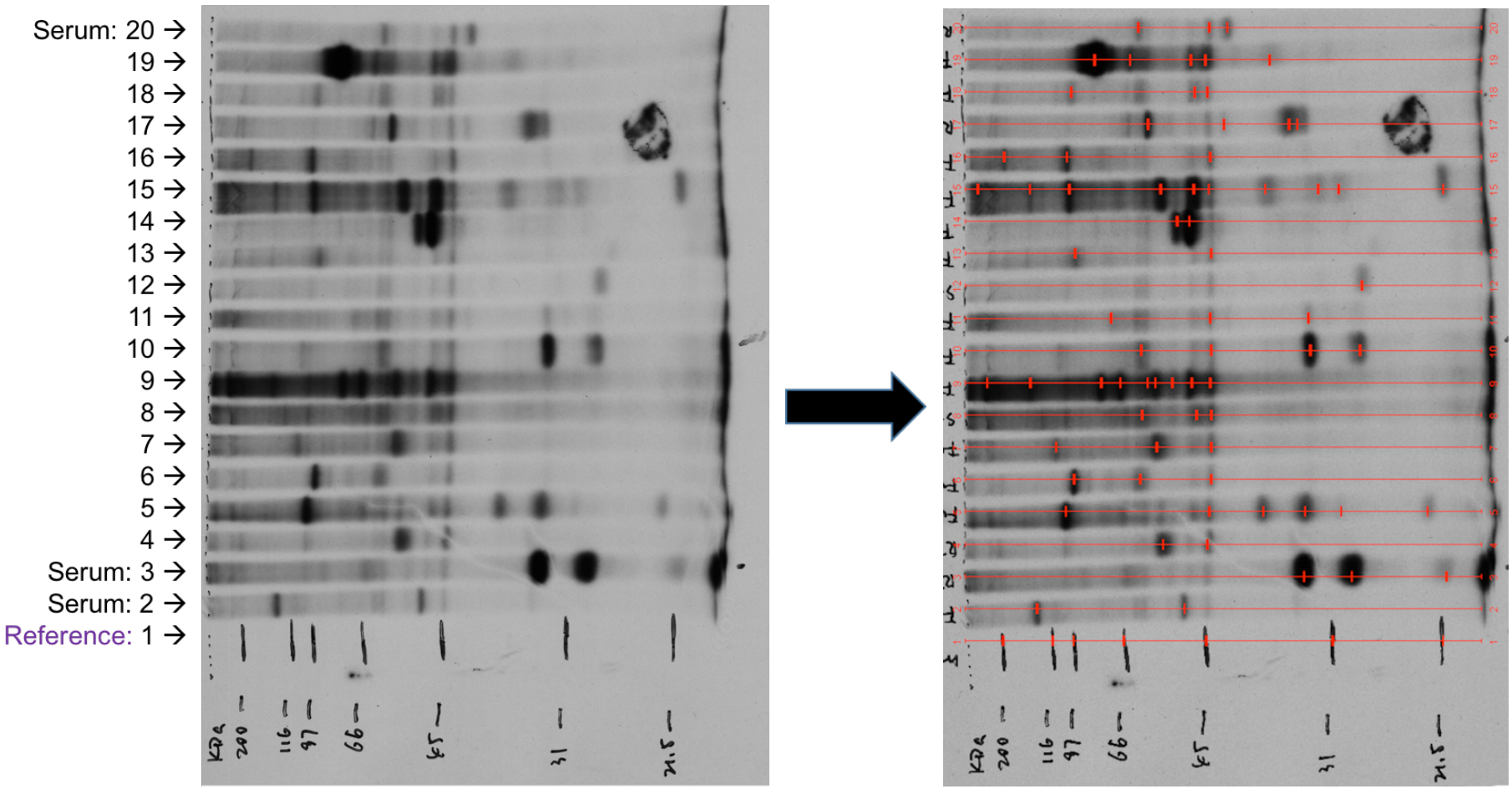


Example II: Raw Data

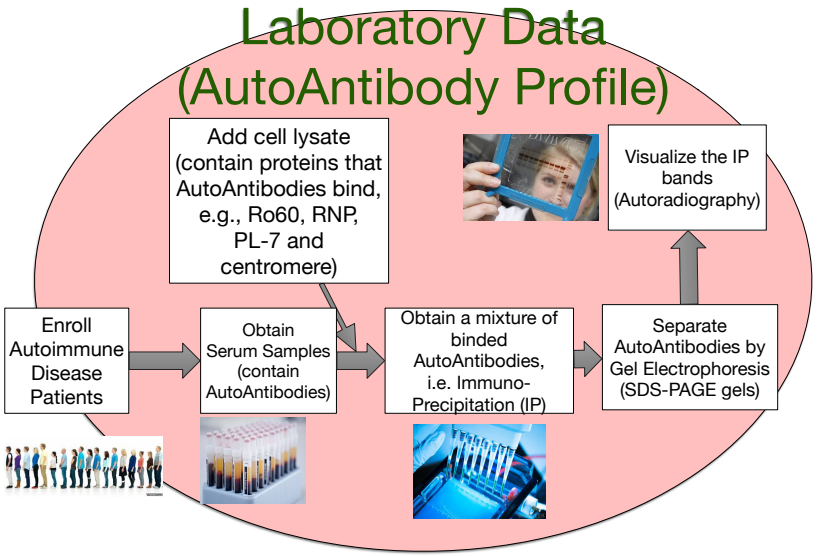
Gel Electrophoresis Autoradiography; 20 Samples

Raw Image

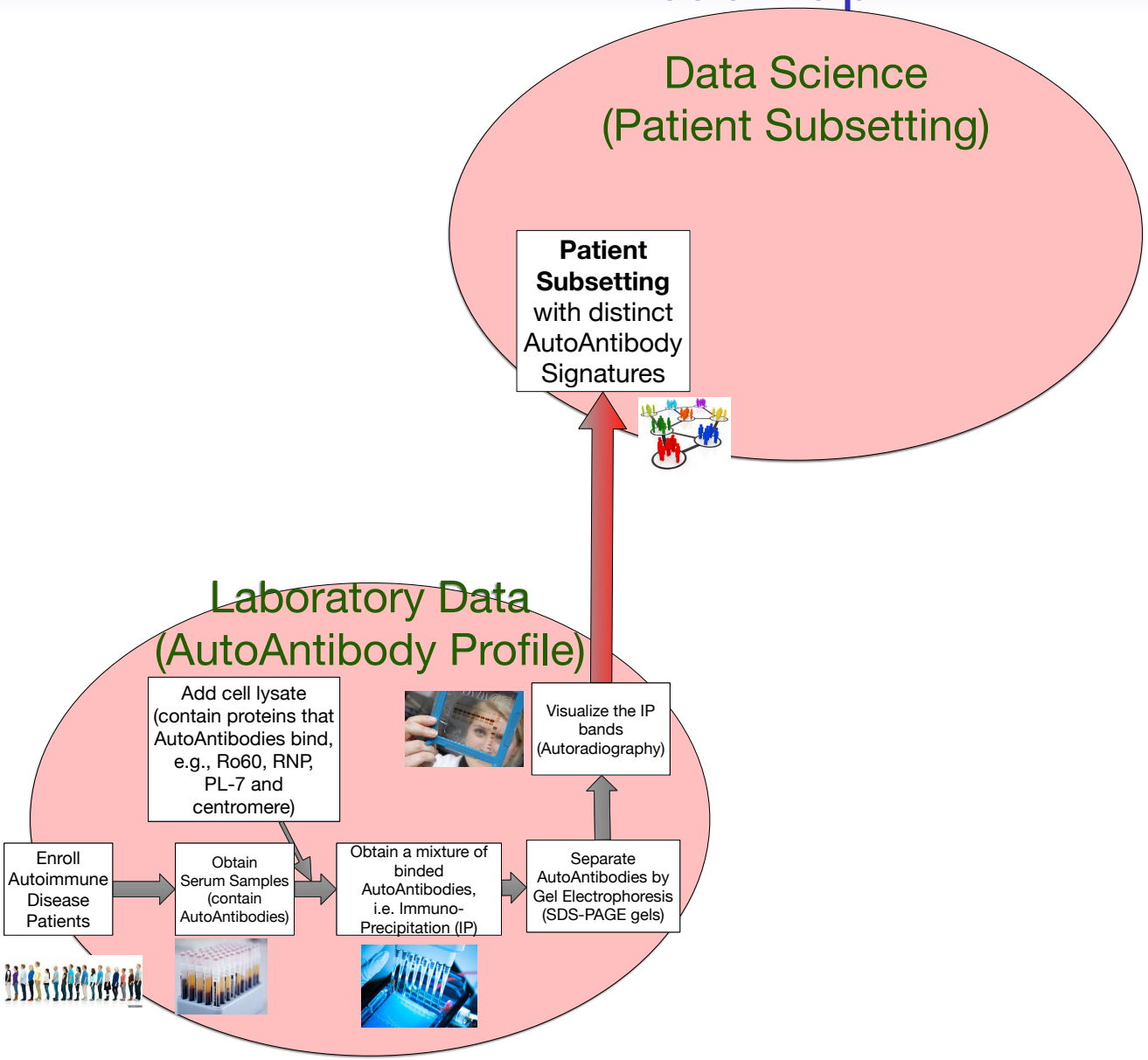
Hand-picked Bands "|"



Roadmap


Laboratory Data
(AutoAntibody Profile)
Add cell lysate
(contain proteins that
AutoAntibodies bind,
e.g., Ro60, RNP,
PL-7 and
centromere)
Visualize the IP
bands
(Autoradiography)
Enroll
Autoimmune
Disease
Patients
Obtain
Serum Samples
(contain
AutoAntibodies)
Obtain a mixture of
binded
AutoAntibodies,
i.e. Immuno-
Precipitation (IP)
Separate
AutoAntibodies by
Gel Electrophoresis
(SDS-PAGE gels)

Roadmap


Data Science
(Patient Subsetting)
Patient Subsetting with distinct AutoAntibody Signatures
Laboratory Data
(AutoAntibody Profile)
Add cell lysate (contain proteins that AutoAntibodies bind, e.g., Ro60, RNP, PL-7 and centromere)
Visualize the IP bands (Autoradiography)
Enroll Autoimmune Disease Patients
Obtain Serum Samples (contain AutoAntibodies)
Obtain a mixture of binded AutoAntibodies, i.e. Immuno-Precipitation (IP)
Separate AutoAntibodies by Gel Electrophoresis (SDS-PAGE gels)

Roadmap

Data Science (Patient Subsetting)

Patient Subsetting with distinct AutoAntibody Signatures

Laboratory Data (AutoAntibody Profile)

Add cell lysate (contain proteins that AutoAntibodies bind, e.g., Ro60, RNP, PL-7 and centromere)

Visualize the IP bands (Autoradiography)

Enroll Autoimmune Disease Patients

Obtain Serum Samples (contain AutoAntibodies)

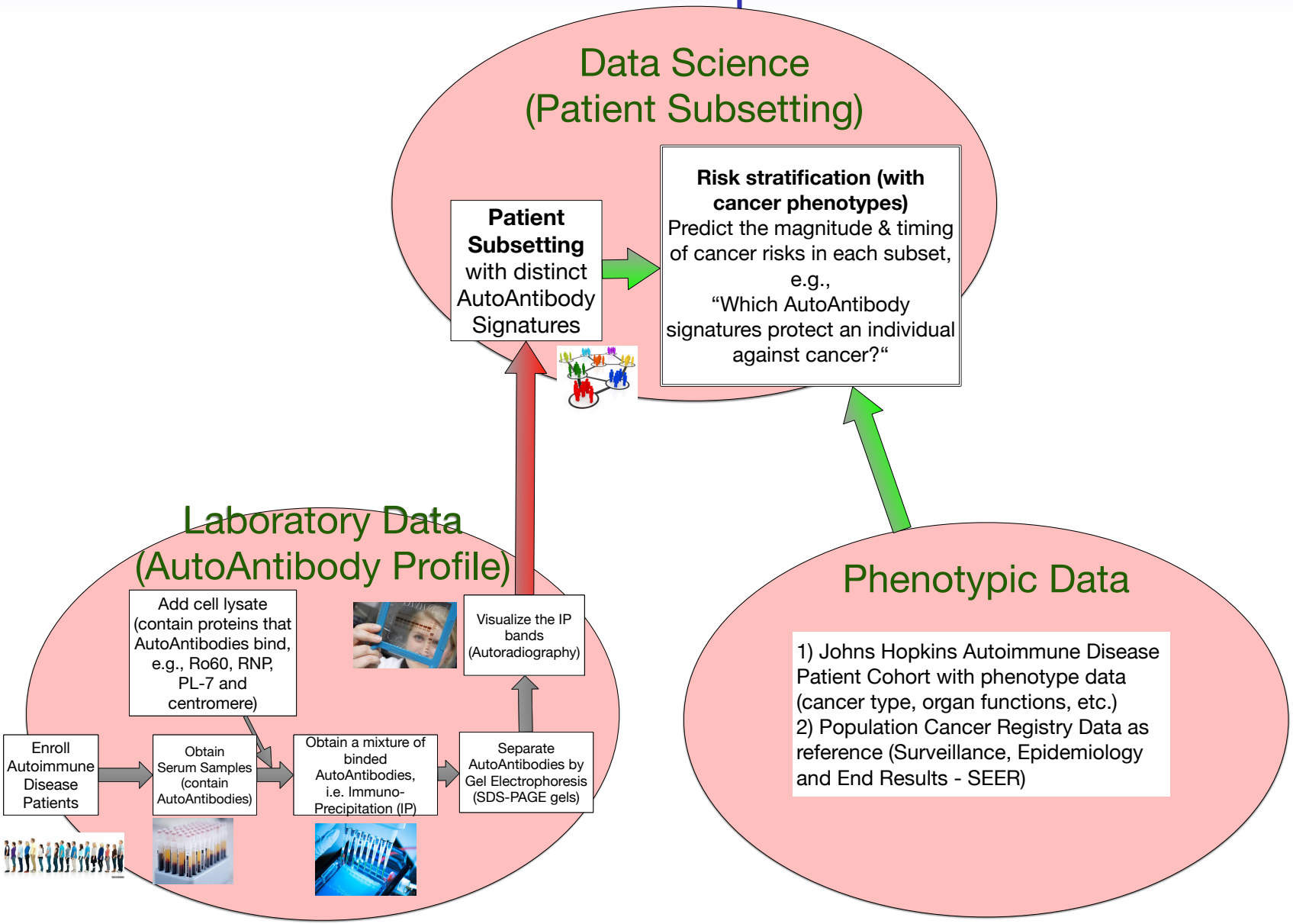
Obtain a mixture of binded AutoAntibodies, i.e. Immuno-Precipitation (IP)

Separate AutoAntibodies by Gel Electrophoresis (SDS-PAGE gels)

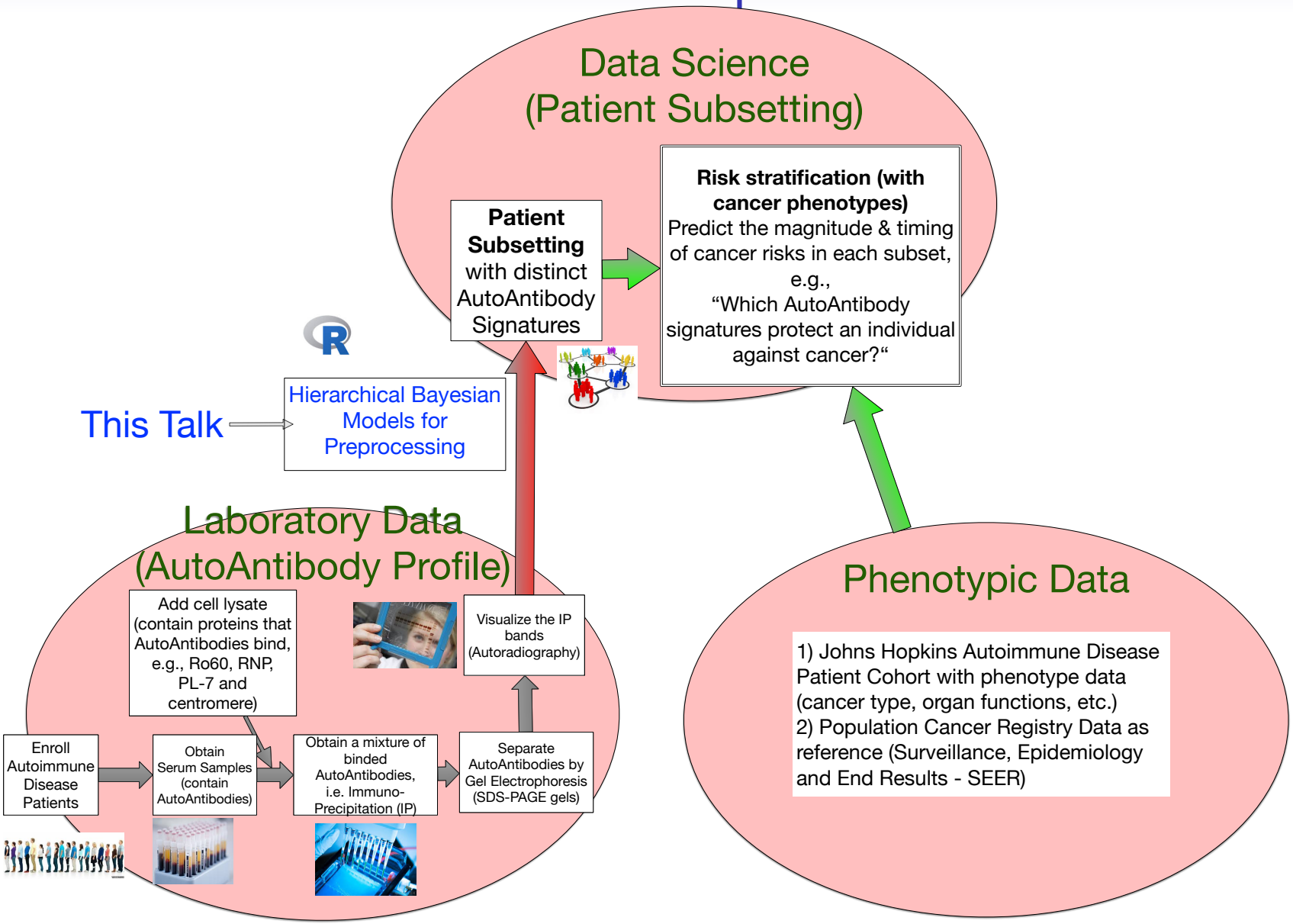
Phenotypic Data

1) Johns Hopkins Autoimmune Disease Patient Cohort with phenotype data (cancer type, organ functions, etc.)
2) Population Cancer Registry Data as reference (Surveillance, Epidemiology and End Results - SEER)

Roadmap


Data Science
(Patient Subsetting)
Patient Subsetting with distinct AutoAntibody Signatures
Risk stratification (with cancer phenotypes)
Predict the magnitude & timing of cancer risks in each subset, e.g.,
"Which AutoAntibody signatures protect an individual against cancer?"
Laboratory Data
(AutoAntibody Profile)
Add cell lysate (contain proteins that AutoAntibodies bind, e.g., Ro60, RNP, PL-7 and centromere)
Visualize the IP bands (Autoradiography)
Enroll Autoimmune Disease Patients
Obtain Serum Samples (contain AutoAntibodies)
Obtain a mixture of binded AutoAntibodies, i.e. Immuno-Precipitation (IP)
Separate AutoAntibodies by Gel Electrophoresis (SDS-PAGE gels)
Phenotypic Data
1) Johns Hopkins Autoimmune Disease Patient Cohort with phenotype data (cancer type, organ functions, etc.)
2) Population Cancer Registry Data as reference (Surveillance, Epidemiology and End Results - SEER)

Roadmap


Data Science
(Patient Subsetting)
Patient Subsetting with distinct AutoAntibody Signatures
Risk stratification (with cancer phenotypes)
Predict the magnitude & timing of cancer risks in each subset, e.g.,
"Which AutoAntibody signatures protect an individual against cancer?"
This Talk
Hierarchical Bayesian Models for Preprocessing
Laboratory Data
(AutoAntibody Profile)
Add cell lysate (contain proteins that AutoAntibodies bind, e.g., Ro60, RNP, PL-7 and centromere)
Visualize the IP bands (Autoradiography)
Enroll Autoimmune Disease Patients
Obtain Serum Samples (contain AutoAntibodies)
Obtain a mixture of binded AutoAntibodies, i.e. Immuno-Precipitation (IP)
Separate AutoAntibodies by Gel Electrophoresis (SDS-PAGE gels)
Phenotypic Data
1) Johns Hopkins Autoimmune Disease Patient Cohort with phenotype data (cancer type, organ functions, etc.)
2) Population Cancer Registry Data as reference (Surveillance, Epidemiology and End Results - SEER)

Autoimmune Diseases and AutoAntibody Signatures

- Etiology of Autoimmue Diseases:

Autoimmune Diseases and AutoAntibody Signatures

- Etiology of Autoimmue Diseases:
  - Human immune system's responses to autoantigens;

Autoimmune Diseases and AutoAntibody Signatures

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

Autoimmune Diseases and AutoAntibody Signatures

- Etiology of Autoimmue 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

Autoimmune Diseases and AutoAntibody Signatures

- Etiology of Autoimmue 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
- Long-term clinical objective: find autoantibody signature that subsets autoimmune disease patients into groups with more homogeneous phenotypes and trajectories

Autoimmune Diseases and AutoAntibody Signatures

- Etiology of Autoimmue 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
- Long-term clinical objective: find autoantibody signature that subsets autoimmune disease patients into groups with more homogeneous phenotypes and trajectories
- Measurements:

Autoimmune Diseases and AutoAntibody Signatures

- Etiology of Autoimmue 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
- Long-term clinical objective: find autoantibody signature that subsets autoimmune disease patients into groups with more homogeneous phenotypes and trajectories
- Measurements: Gel Electrophoresis Autoradiography (GEA)

Gel Electrophoresis Autoradiography (GEA)

A technique to visualize the abundance of molecules or fragments of molecules that have been radioactively labeled.

- Can generate 100s of possibilities of band patterns

Gel Electrophoresis Autoradiography (GEA)

A technique to visualize the abundance of molecules or fragments of molecules that have been radioactively labeled.

- Can generate 100s of possibilities of band patterns
- Can be tested and validated using commercially available line immunoblot assay (EuroImmun; Systemic Sclerosis (Nucleoli) profile)

Gel Electrophoresis Autoradiography (GEA)

A technique to visualize the abundance of molecules or fragments of molecules that have been radioactively labeled.

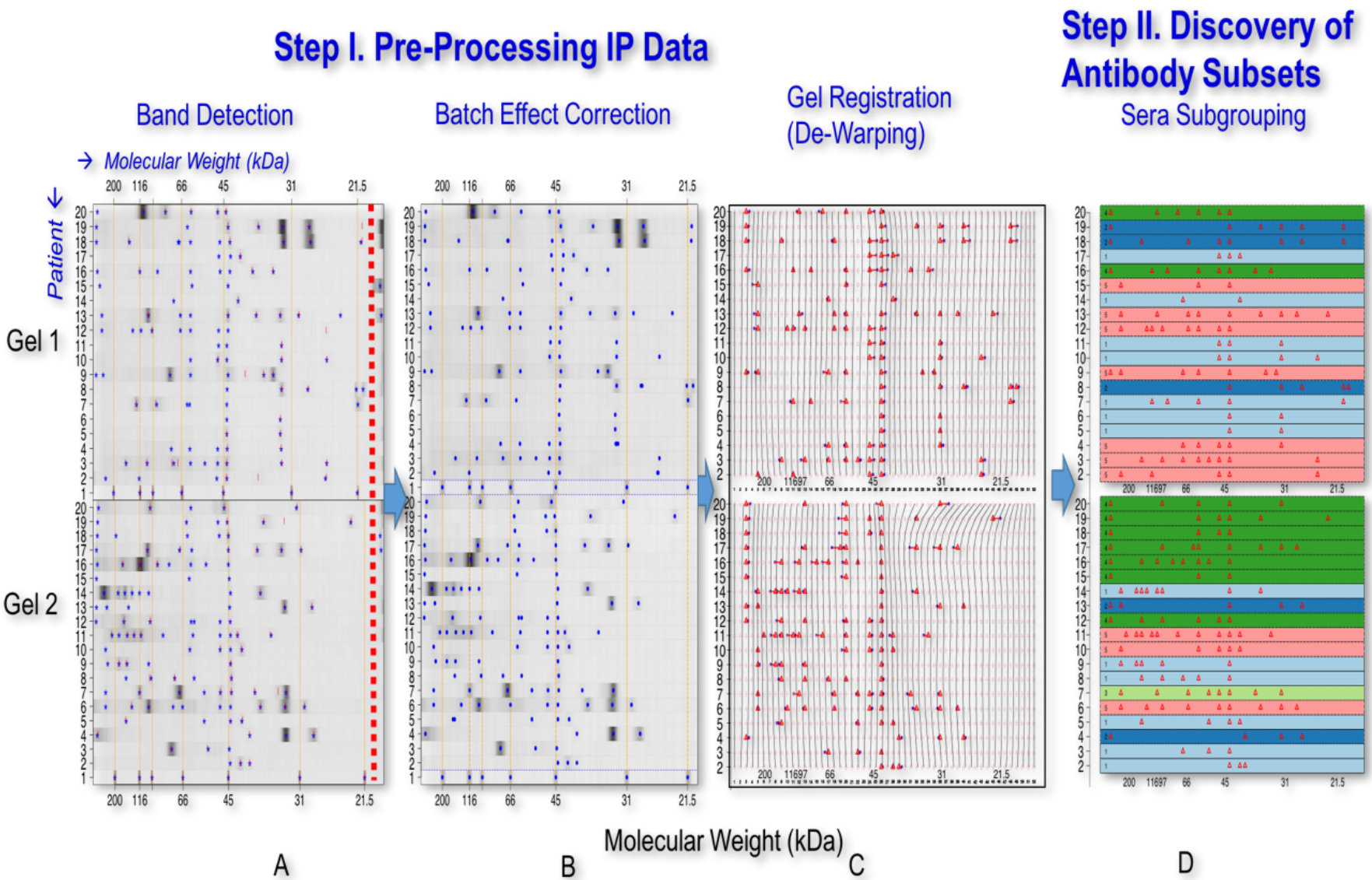
- Can generate 100s of possibilities of band patterns
- Can be tested and validated using commercially available line immunoblot assay (EuroImmun; Systemic Sclerosis (Nucleoli) profile)
- Gap: Onerous and expensive to validate; Need a method to greatly simplify autoantibody profile discovery

Gel Electrophoresis Autoradiography (GEA)

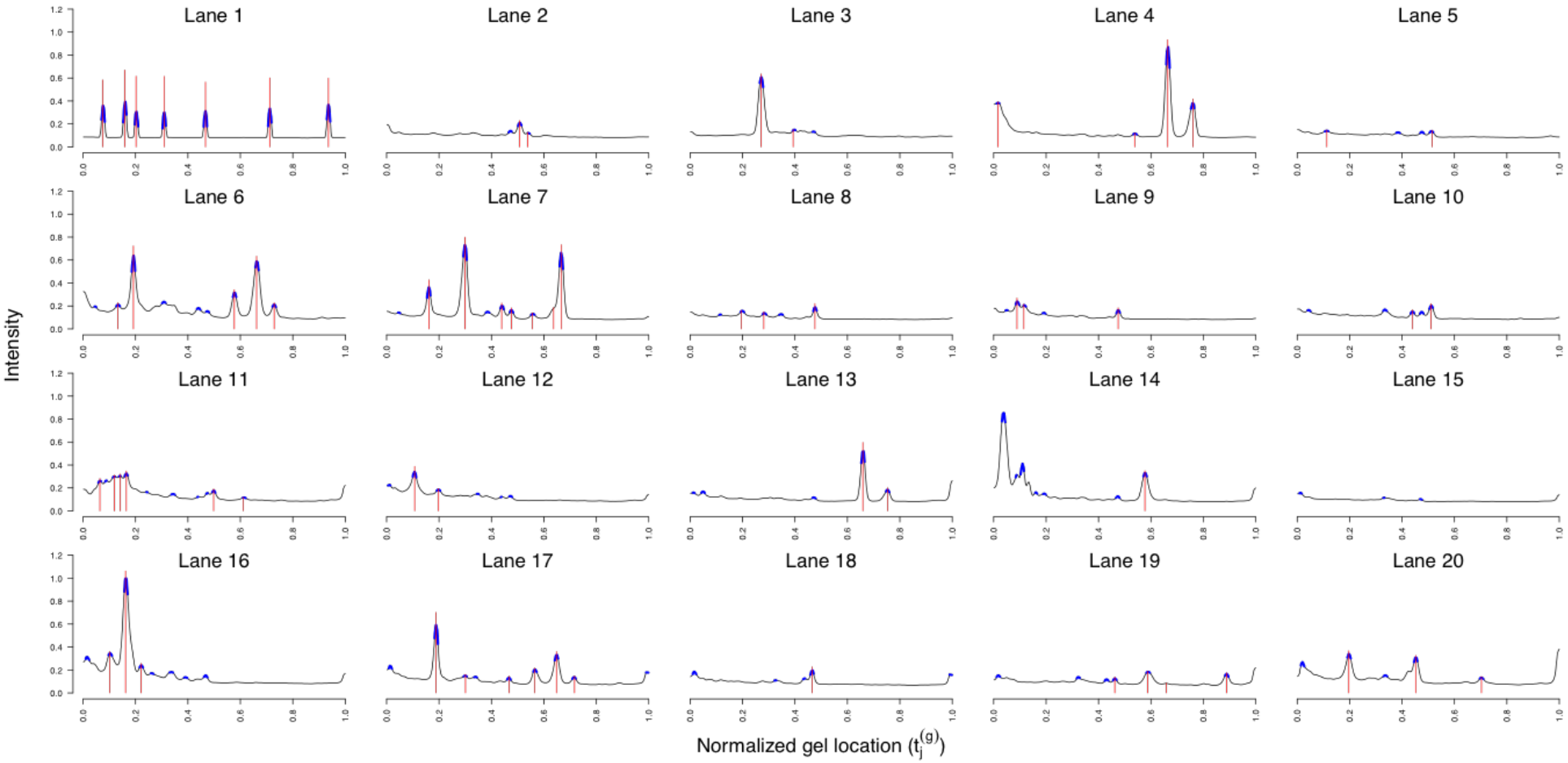
A technique to visualize the abundance of molecules or fragments of molecules that have been radioactively labeled.

- Can generate 100s of possibilities of band patterns
- Can be tested and validated using commercially available line immunoblot assay (EuroImmun; Systemic Sclerosis (Nucleoli) profile)
- Gap: Onerous and expensive to validate; Need a method to greatly simplify autoantibody profile discovery
- Solution: Pre-filtering to define subgroups with similar specificities based on the bands observed by GEA

Automated Pipeline for Autoimmune Disease Subsetting

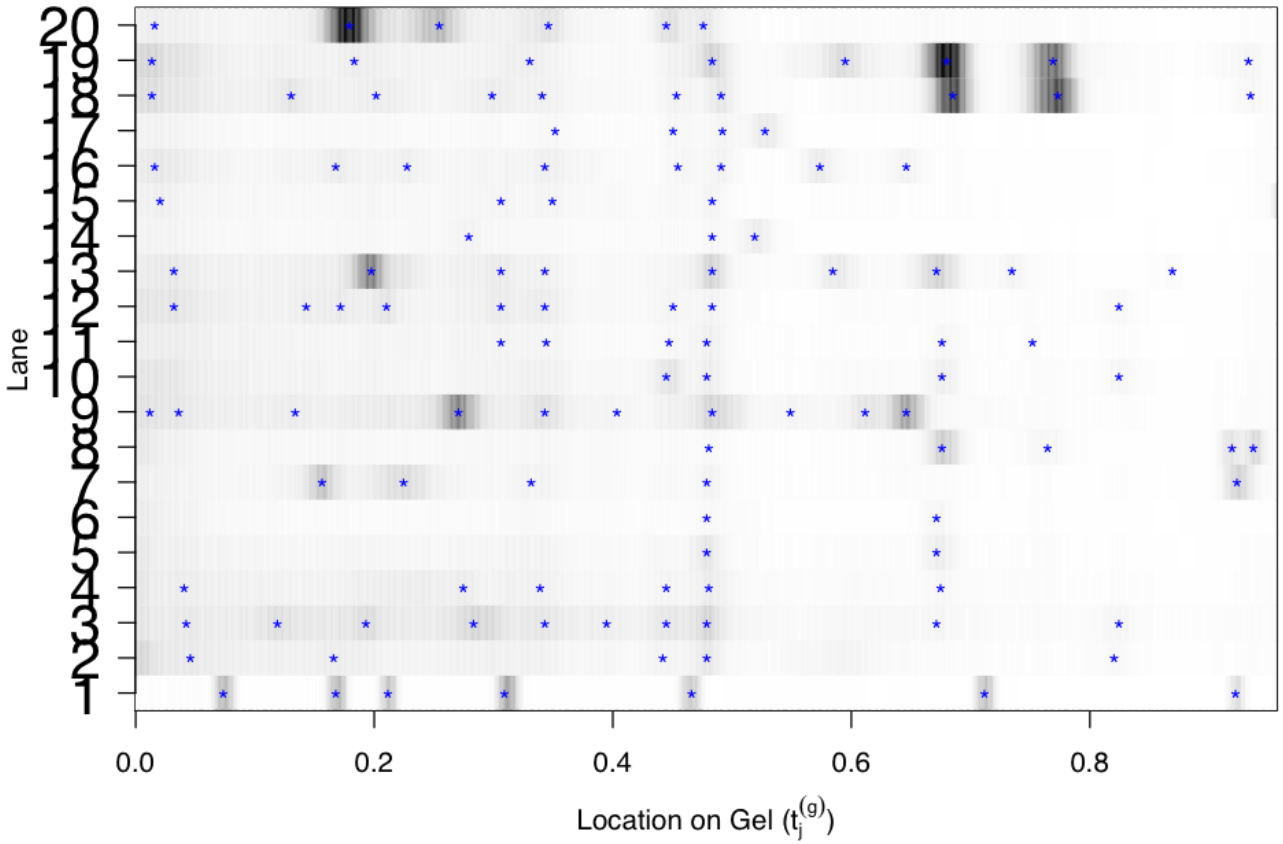


Step I-A: Automated Peak Detection



Step I-A: Automated Peak Detection

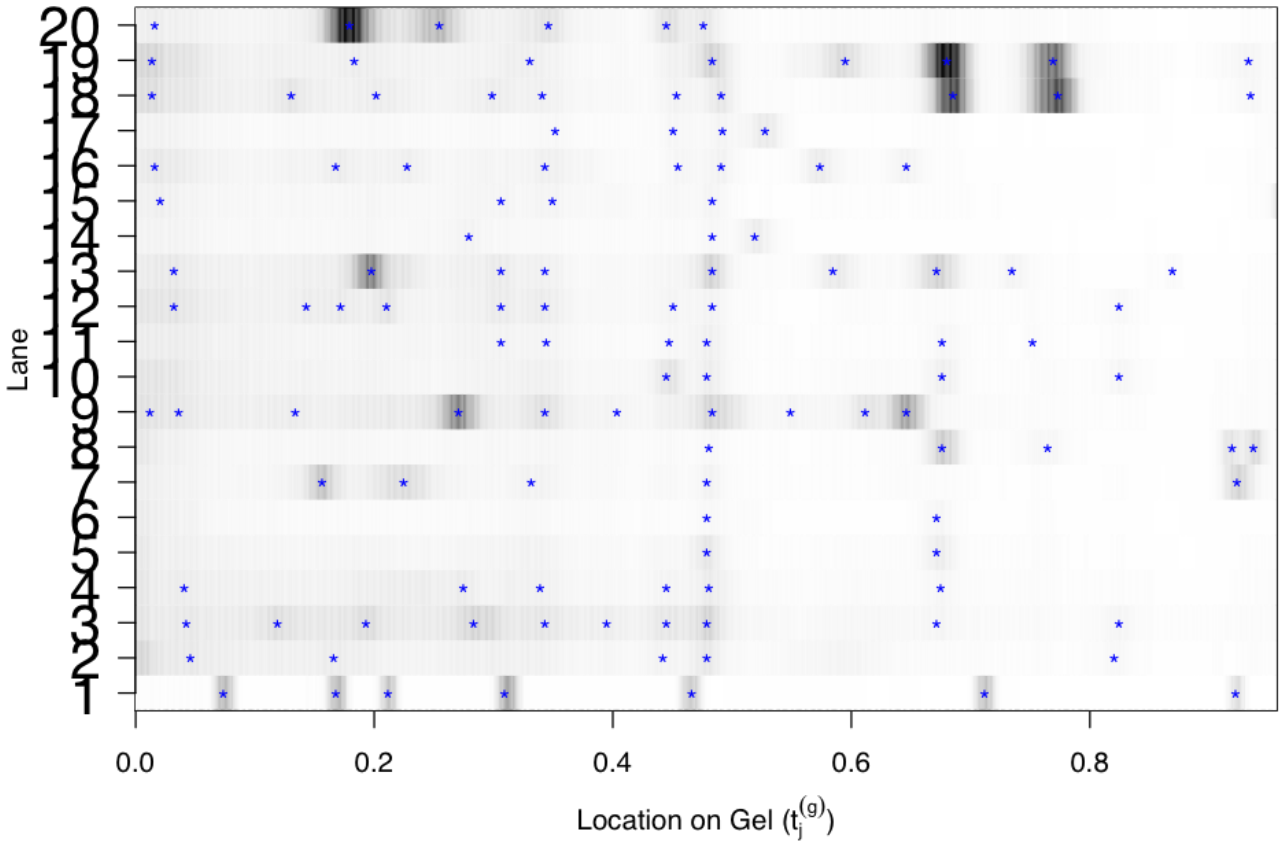
Overlaid against gel image; "$*$" for detected peaks



- u_{gi}: lane number for lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$

Step I-A: Automated Peak Detection

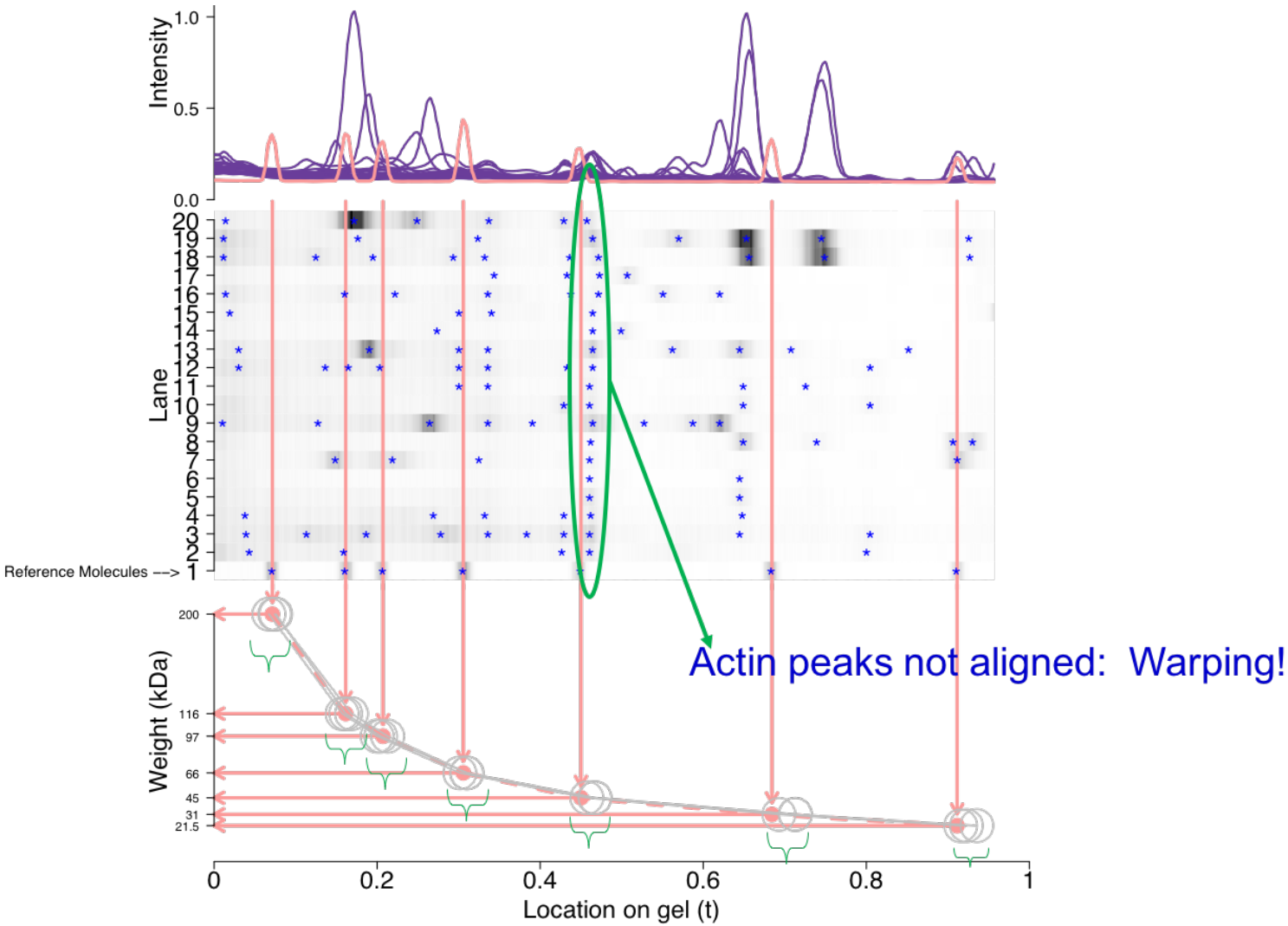
Overlaid against gel image; "$*$" for detected peaks



- u_{gi}: lane number for lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$
- T_{gij}: location for the j-th peak ("$*$"), $j = 1, \ldots, J_{gi}$, for lane i, gel g

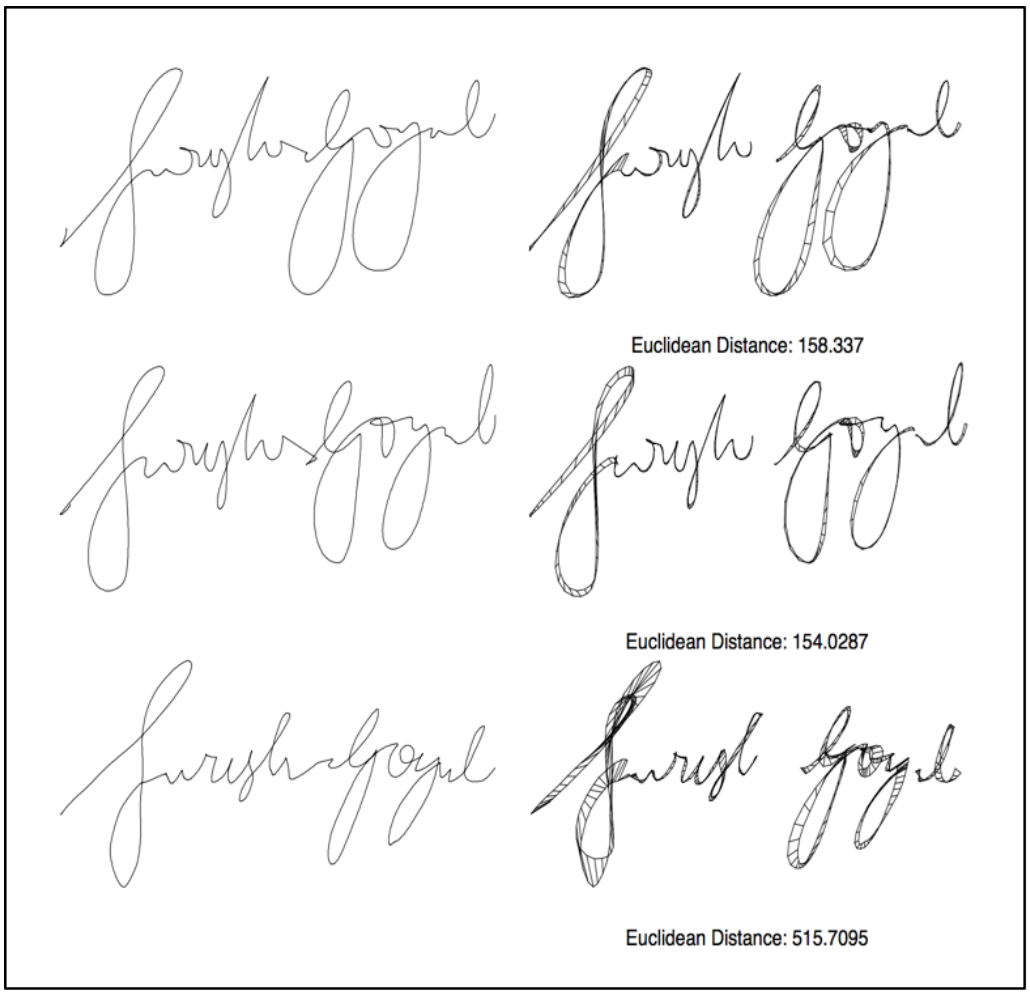
Step I-B: Batch Effect Correction

Must address before meaningful subgrouping

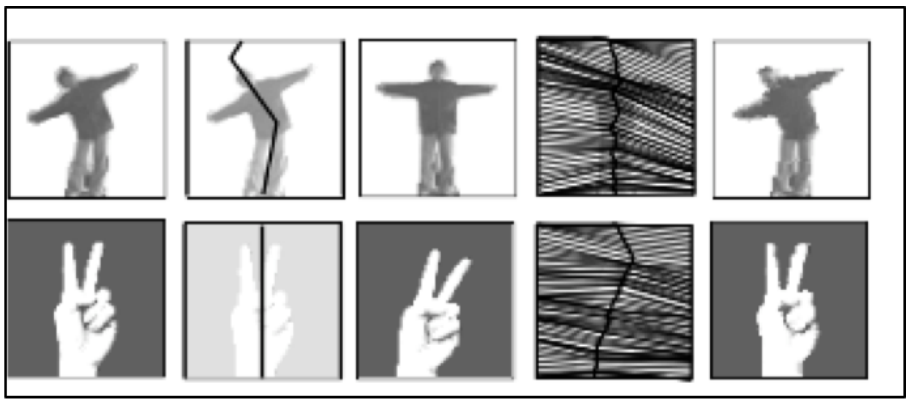


Reference proteins not aligned: Batch effects!

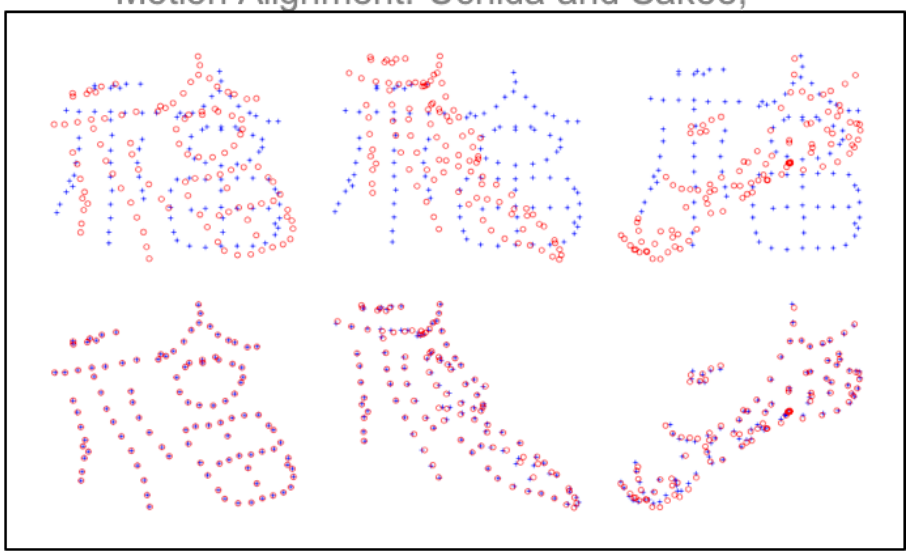
Warping Examples



Signature Verification: Hastie et al. 1991



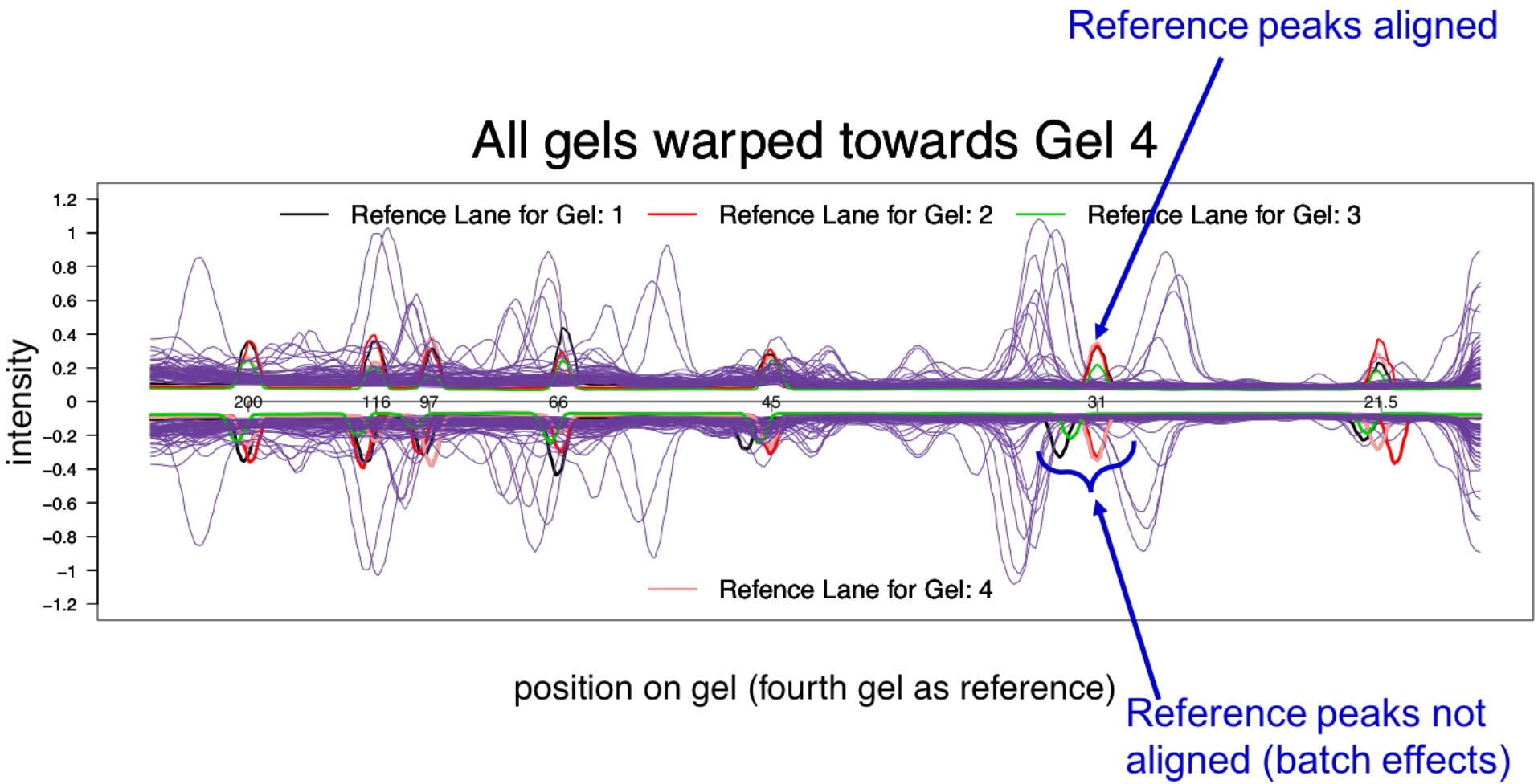
Motion Alignment: Uchida and Sakoe,



Handwritten Chinese: Ma and Zhao (2015)

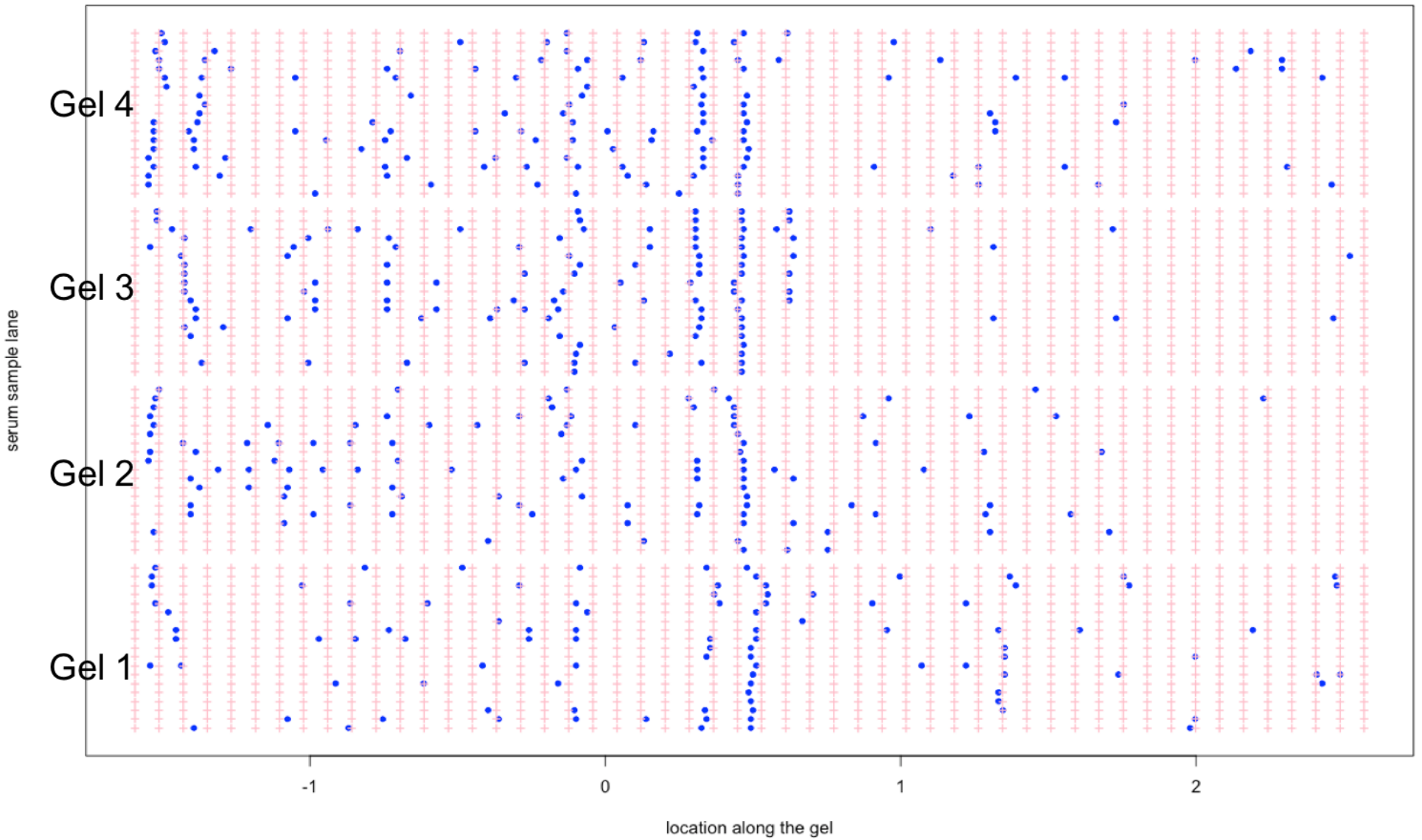
Step I-B: Batch Effect Correction

Piecewise Linear Warping by Reference Marker Molecules



Step I-C: Align the peaks

"Which "+" do the peaks "•" belong?"



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 assume smoothness of 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 assume smoothness of warping
  - Ubiquitous proteins (e.g., actin) on multiple gels must be aligned.
    Solution: Discretized non-homogeneous Poisson process with shared intensity across gels

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 assume smoothness of warping
  - Ubiquitous proteins (e.g., actin) on multiple gels must be aligned.
    Solution: Discretized non-homogeneous Poisson process with shared intensity across gels
  - The observed peak locations are noisy.
    Solution: Gaussian noise around the true location

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

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

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:
  1. $Z_{gij} \in \{1, \ldots, L\}$, $j = 1, \ldots, J_{gi}$ (match a "$*$" to a "$+$"), 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

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:
  1. $Z_{gij} \in \{1, \ldots, L\}$, $j = 1, \ldots, J_{gi}$ (match a "$*$" to a "$+$"), 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

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:
  1. $Z_{gij} \in \{1, \ldots, L\}$, $j = 1, \ldots, J_{gi}$ (match a "$*$" to a "$+$"), 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
  - Number of observed peaks in lane i, gel g: $J_{gi} \overset{d}{\sim} \text{Poisson}(\Lambda)$

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:
  1. $Z_{gij} \in \{1, \ldots, L\}$, $j = 1, \ldots, J_{gi}$ (match a "$*$" to a "$+$"), 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
  - Number of observed peaks in lane i, gel g:
    $J_{gi} \overset{d}{\sim} \text{Poisson}(\Lambda)$
    Λ: Cumulative intensity; Controls the total number of peaks

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:
  1. $Z_{gij} \in \{1, \ldots, L\}$, $j = 1, \ldots, J_{gi}$ (match a "$*$" to a "$+$"), 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
  - Number of observed peaks in lane i, gel g:
    $J_{gi} \overset{d}{\sim} \text{Poisson}(\Lambda)$
    Λ: Cumulative intensity; Controls the total number of peaks
  - Peak-to-landmark indicators:
    $(Z_{gi1}, \ldots, Z_{giJ_{gi}}) =$ increasing sort $\{Z^*_{gi1}, \ldots, Z^*_{giJ_{gi}}\}$

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:
  1. $Z_{gij} \in \{1, \ldots, L\}$, $j = 1, \ldots, J_{gi}$ (match a "$*$" to a "$+$"), 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
  - Number of observed peaks in lane i, gel g:
    $J_{gi} \overset{d}{\sim}$ Poisson(Λ)
    Λ: Cumulative intensity; Controls the total number of peaks
  - Peak-to-landmark indicators:
    $(Z_{gi1}, \ldots, Z_{giJ_{gi}}) =$ increasing sort $\{Z^*_{gi1}, \ldots, Z^*_{giJ_{gi}}\}$
  - $Z^*_{gij} \overset{iid}{\sim}$ Categorical $\left(\{\lambda^*_\ell\}_{\ell=1}^{L}\right)$

Step I-C: Model for 2-Dimensional Image Dewarping

Prior on the peak-to-landmark indicators

- Peak-to-landmark Indicators:
  1. $Z_{gij} \in \{1, \ldots, L\}$, $j = 1, \ldots, J_{gi}$ (match a "$*$" to a "$+$"), 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
  - Number of observed peaks in lane i, gel g:
    $J_{gi} \overset{d}{\sim}$ Poisson(Λ)
    Λ: Cumulative intensity; Controls the total number of peaks
  - Peak-to-landmark indicators:
    $(Z_{gi1}, \ldots, Z_{giJ_{gi}}) =$ increasing sort $\{Z^*_{gi1}, \ldots, Z^*_{giJ_{gi}}\}$
  - $Z^*_{gij} \overset{iid}{\sim}$ Categorical $\left(\{\lambda^*_\ell\}_{\ell=1}^{L}\right)$
    λ^*_ℓ: 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 Dewarping

Gaussian Mixture Model for Noisy Peak Locations "$*$"

- Model the observed peaks T_{gij} as observations from a L-component Gaussian mixture, for each candidate landmark ℓ

Step I-C: 2-Dimensional Image Dewarping

Gaussian Mixture Model for Noisy Peak Locations "$*$"

- Model the observed peaks T_{gij} as observations from a L-component Gaussian mixture, for each candidate landmark ℓ
- We assume

$$p\left\{ (\underbrace{T_{gij}=t}_{\substack{\text{peak}\\ \text{location}}}, \underbrace{u_{gi}}_{\substack{\text{lane}\\ \text{number}}}) \mid \underbrace{Z_{gij}=\ell}_{\substack{\text{matched to}\\ \text{landmark } \ell}}, \underbrace{T_{gi,j-1}}_{\substack{\text{nearest left}\\ \text{peak location}}}, \underbrace{\mathcal{S}_g}_{\substack{\text{warping}\\ \text{function}}}, \underbrace{\sigma_\epsilon}_{\substack{\text{noise}\\ \text{level}}} \right\}$$

$$= \begin{cases} \phi\left(t; \mathcal{S}_g(\nu_\ell, u_{gi}), \sigma_\epsilon\right), & t \in \mathcal{I}_{gij}(\nu_\ell, A_0); \\ 0, & \text{otherwise}, \end{cases}$$

$\ell = 1, \ldots, L$, peak $j = 1, \ldots, J_{gi}$, lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$.

Step I-C: 2-Dimensional Image Dewarping

Gaussian Mixture Model for Noisy Peak Locations "$*$"

- Model the observed peaks T_{gij} as observations from a L-component Gaussian mixture, for each candidate landmark ℓ
- We assume

$$p\left\{ \left(\underbrace{T_{gij}=t}_{\text{peak location}},\ \underbrace{u_{gi}}_{\text{lane number}}\right) \mid \underbrace{Z_{gij}=\ell}_{\text{matched to landmark } \ell},\ \underbrace{T_{gi,j-1}}_{\text{nearest left peak location}},\ \underbrace{\mathcal{S}_g}_{\text{warping function}},\ \underbrace{\sigma_\epsilon}_{\text{noise level}} \right\}$$

$$= \begin{cases} \phi\left(t; \mathcal{S}_g(\nu_\ell, u_{gi}), \sigma_\epsilon\right), & t \in \mathcal{I}_{gij}(\nu_\ell, A_0); \\ 0, & \text{otherwise}, \end{cases}$$

$\ell = 1, \ldots, L$, peak $j = 1, \ldots, J_{gi}$, lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$.

- $\phi(\cdot; a, b)$: Gaussian density with mean a and standard deviation b.

Step I-C: 2-Dimensional Image Dewarping

Gaussian Mixture Model for Noisy Peak Locations "$*$"

- Model the observed peaks T_{gij} as observations from a L-component Gaussian mixture, for each candidate landmark ℓ
- We assume

$$p\left\{ (\underbrace{T_{gij}=t}_{\substack{\text{peak}\\\text{location}}},\ \underbrace{u_{gi}}_{\substack{\text{lane}\\\text{number}}}) \mid \underbrace{Z_{gij}=\ell}_{\substack{\text{matched to}\\\text{landmark } \ell}},\ \underbrace{T_{gi,j-1}}_{\substack{\text{nearest left}\\\text{peak location}}},\ \underbrace{\mathcal{S}_g}_{\substack{\text{warping}\\\text{function}}},\ \underbrace{\sigma_\epsilon}_{\substack{\text{noise}\\\text{level}}} \right\}$$

$$= \begin{cases} \phi\left(t; \mathcal{S}_g(\nu_\ell, u_{gi}), \sigma_\epsilon\right), & t \in \mathcal{I}_{gij}(\nu_\ell, A_0); \\ 0, & \text{otherwise}, \end{cases}$$

$\ell = 1, \ldots, L$, peak $j = 1, \ldots, J_{gi}$, lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$.

- $\mathcal{S}_g$: $(\nu_\ell, u_{gi}) \mapsto \mathcal{S}_g(\nu_\ell, u_i)$, *unknown*, smooth bivariate function for the spatial deformation

Step I-C: 2-Dimensional Image Dewarping

Gaussian Mixture Model for Noisy Peak Locations "$*$"

- Model the observed peaks T_{gij} as observations from a L-component Gaussian mixture, for each candidate landmark ℓ
- We assume

$$p\left\{ (\underbrace{T_{gij}=t}_{\substack{\text{peak}\\ \text{location}}}, \underbrace{u_{gi}}_{\substack{\text{lane}\\ \text{number}}}) \mid \underbrace{Z_{gij}=\ell}_{\substack{\text{matched to}\\ \text{landmark } \ell}}, \underbrace{T_{gi,j-1}}_{\substack{\text{nearest left}\\ \text{peak location}}}, \underbrace{\mathcal{S}_g}_{\substack{\text{warping}\\ \text{function}}}, \underbrace{\sigma_\epsilon}_{\substack{\text{noise}\\ \text{level}}} \right\}$$

$$= \begin{cases} \phi\left(t; \mathcal{S}_g(\nu_\ell, u_{gi}), \sigma_\epsilon\right), & t \in \mathcal{I}_{gij}(\nu_\ell, A_0); \\ 0, & \text{otherwise,} \end{cases}$$

$\ell = 1, \ldots, L$, peak $j = 1, \ldots, J_{gi}$, lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$.

- 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

Step I-C: 2-Dimensional Image Dewarping

Gaussian Mixture Model for Noisy Peak Locations "$*$"

- Model the observed peaks T_{gij} as observations from a L-component Gaussian mixture, for each candidate landmark ℓ
- We assume

$$p\left\{ \left(\underbrace{T_{gij}=t}_{\text{peak location}},\ \underbrace{u_{gi}}_{\text{lane number}}\right) \mid \underbrace{Z_{gij}=\ell}_{\text{matched to landmark } \ell},\ \underbrace{T_{gi,j-1}}_{\text{nearest left peak location}},\ \underbrace{\mathcal{S}_g}_{\text{warping function}},\ \underbrace{\sigma_\epsilon}_{\text{noise level}} \right\}$$

$$= \begin{cases} \phi\left(t; \mathcal{S}_g(\nu_\ell, u_{gi}), \sigma_\epsilon\right), & t \in \mathcal{I}_{gij}(\nu_\ell, A_0); \\ 0, & \text{otherwise}, \end{cases}$$

$\ell = 1, \ldots, L$, peak $j = 1, \ldots, J_{gi}$, lane $i = 1, \ldots, N_g$, gel $g = 1, \ldots, G$.

- Let $\mathcal{P}_g$ be the peaks for gel g; let $\mathcal{P}$ collect all the peaks

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively
  - $\{\beta_{gst}\}$: the set of coefficients to be estimated

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively
  - $\{\beta_{gst}\}$: the set of coefficients to be estimated
- Implementing Warping Function Constraints and Priors

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively
  - $\{\beta_{gst}\}$: the set of coefficients to be estimated
- Implementing Warping Function Constraints and Priors
  - Boundary constraint: $\mathcal{S}_g(\nu_0, u) = \nu_0, \mathcal{S}_g(\nu_{L+1}, u) = \nu_{L+1}$

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively
  - $\{\beta_{gst}\}$: the set of coefficients to be estimated
- Implementing Warping Function Constraints and Priors
  - Boundary constraint: $\mathcal{S}_g(\nu_0, u) = \nu_0, \mathcal{S}_g(\nu_{L+1}, u) = \nu_{L+1}$
  - Monotonic constraint:
    $\nu_0 \leq \mathcal{S}_g(\nu, u) < \mathcal{S}_g(\nu', u \leq \nu_{L+1}, \forall \nu < \nu', \forall u$

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively
  - $\{\beta_{gst}\}$: the set of coefficients to be estimated
- Implementing Warping Function Constraints and Priors
  - Boundary constraint: $\mathcal{S}_g(\nu_0, u) = \nu_0, \mathcal{S}_g(\nu_{L+1}, u) = \nu_{L+1}$
  - Monotonic constraint: $\nu_0 \leq \mathcal{S}_g(\nu, u) < \mathcal{S}_g(\nu', u \leq \nu_{L+1}, \forall \nu < \nu', \forall u$
  - Both constraints above can be implemented via constraints on $\{\beta_{gst}\}$

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively
  - $\{\beta_{gst}\}$: the set of coefficients to be estimated
- Implementing Warping Function Constraints and Priors
  - Boundary constraint: $\mathcal{S}_g(\nu_0, u) = \nu_0, \mathcal{S}_g(\nu_{L+1}, u) = \nu_{L+1}$
  - Monotonic constraint: $\nu_0 \leq \mathcal{S}_g(\nu, u) < \mathcal{S}_g(\nu', u \leq \nu_{L+1}, \forall \nu < \nu', \forall u$
  - Both constraints above can be implemented via constraints on $\{\beta_{gst}\}$
  - Smoothness: Bayesian penalized-splines to make adjacent $\{\beta_{gst}\}$ similar

Step I-C: 2-Dimensional Image Dewarping

Warping Function by Tensor Product Basis Expansion

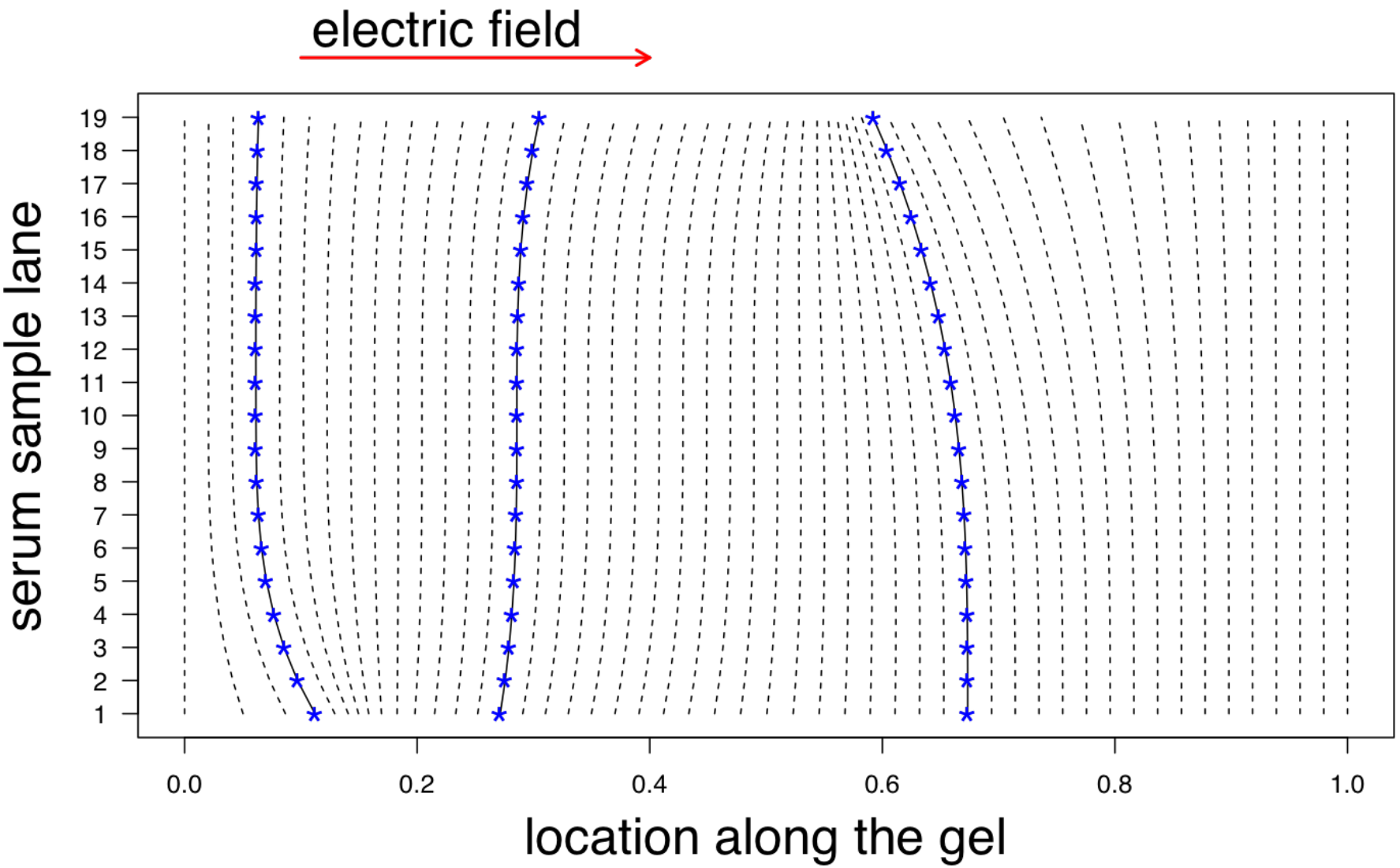
- We assume the warping function

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

  - $B_{g1s}(\cdot)$ and $B_{g2t}(\cdot)$: the s-th and t-th cubic B-spline basis along the two coordinate directions, respectively
  - $\{\beta_{gst}\}$: the set of coefficients to be estimated
- Implementing Warping Function Constraints and Priors
  - Boundary constraint: $\mathcal{S}_g(\nu_0, u) = \nu_0, \mathcal{S}_g(\nu_{L+1}, u) = \nu_{L+1}$
  - Monotonic constraint: $\nu_0 \leq \mathcal{S}_g(\nu, u) < \mathcal{S}_g(\nu', u \leq \nu_{L+1}, \forall \nu < \nu', \forall u$
  - Both constraints above can be implemented via constraints on $\{\beta_{gst}\}$
  - Smoothness: Bayesian penalized-splines to make adjacent $\{\beta_{gst}\}$ similar
  - Vary by gel: $\mathcal{S}_g(\nu_\ell, u)$

Step I-C: A Mathematical Model for Warping

Estimate the warping, then reverse



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

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

Recall:

- $\mathbf{Z}$: the collection of peak-to-landmark indicators

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

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

Recall:

- $\mathbf{Z}$: the collection of peak-to-landmark indicators
- $\mathcal{P}$: the collection of all the observed peaks

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

Joint distribution $[\mathcal{P}, \mathbf{Z}]$:

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

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

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

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

- Goal: Joint distribution $[\mathcal{P}, \mathbf{Z}]$(data+unknowns) $\rightarrow$ Posterior distribution $[\mathbf{Z} \mid \mathcal{P}]$ (unknown given data)

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

Joint distribution $[\mathcal{P}, \mathbf{Z}]$:

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

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

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

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

- Goal: Joint distribution $[\mathcal{P}, \mathbf{Z}]$(data+unknowns) $\rightarrow$ Posterior distribution $[\mathbf{Z} \mid \mathcal{P}]$ (unknown given data)
- Tool: Markov chain Monte Carlo (MCMC)

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

Joint distribution $[\mathcal{P}, \mathbf{Z}]$:

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

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

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

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

- Goal: Joint distribution $[\mathcal{P}, \mathbf{Z}]$(data+unknowns) $\rightarrow$ Posterior distribution $[\mathbf{Z} \mid \mathcal{P}]$ (unknown given data)
- Tool: Markov chain Monte Carlo (MCMC)
- Idea: Simulate samples from the joint posterior distribution of the unknowns given the data;

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

Joint distribution $[\mathcal{P}, \mathbf{Z}]$:

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

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

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

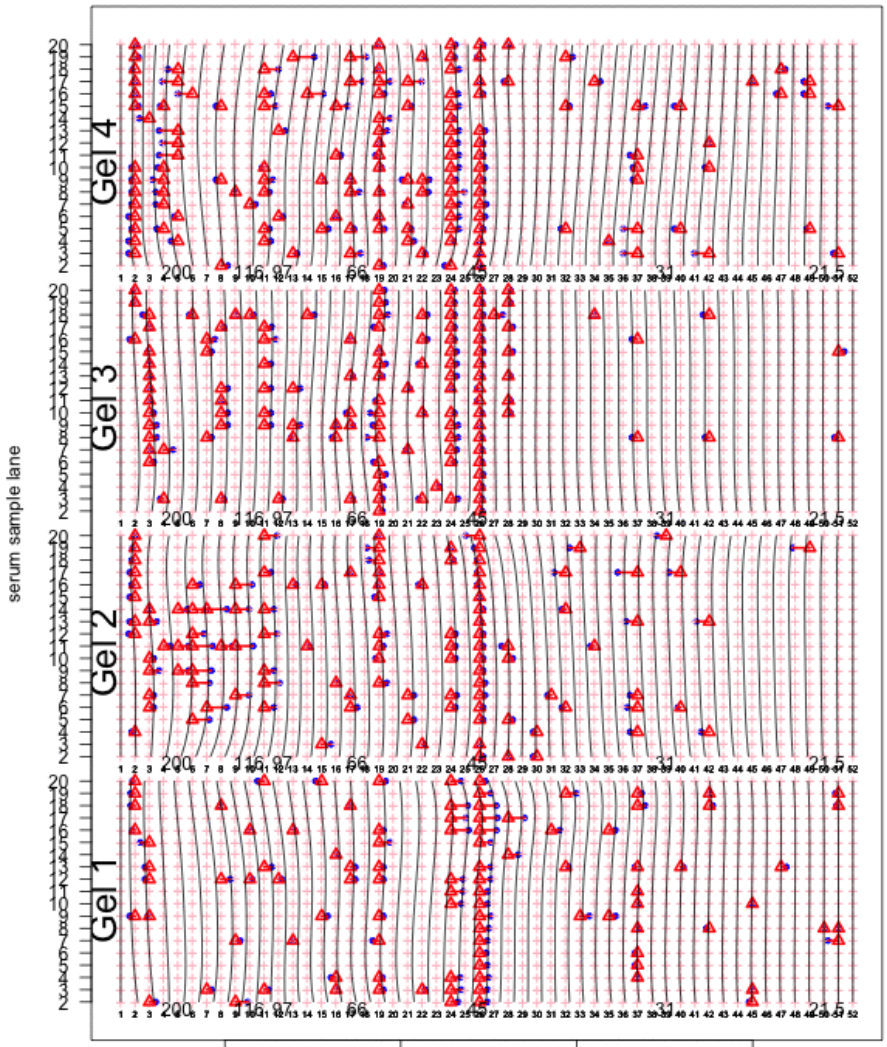
$$\times\underbrace{\prod_{s=2}^{T_\nu-1}\left[N_{T_u}\left(\{\beta_{gst}\}_{t=1}^{T_u}; \mathbf{0}, \sigma_{gs}^{-2}\Delta_2'\Delta_2\right)\cdot p(\sigma_{gs}^2,\rho_g)\right]\Bigg\}}_{\text{prior (2.7) and hyperpriors of the smoothing parameters}}\times\underbrace{p(\boldsymbol{\lambda})}_{\text{hyperprior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

- Goal: Joint distribution $[\mathcal{P}, \mathbf{Z}]$(data+unknowns) $\rightarrow$ Posterior distribution $[\mathbf{Z} \mid \mathcal{P}]$ (unknown given data)
- Tool: Markov chain Monte Carlo (MCMC)
- Idea: Simulate samples from the joint posterior distribution of the unknowns given the data; Then use the samples to do posterior inference for any functions of the unknowns

Step I-C: Align the peaks – Result

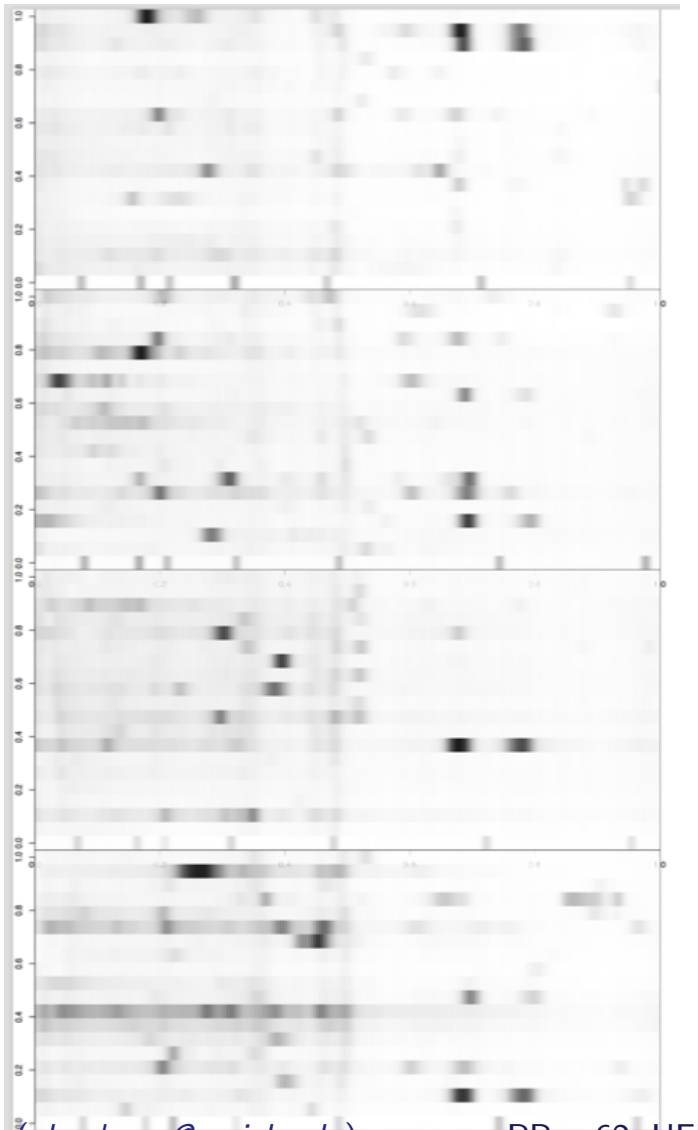
Animation; "Δ" for signature; "•" for the observed peaks

(Please Click the Image for Animation)



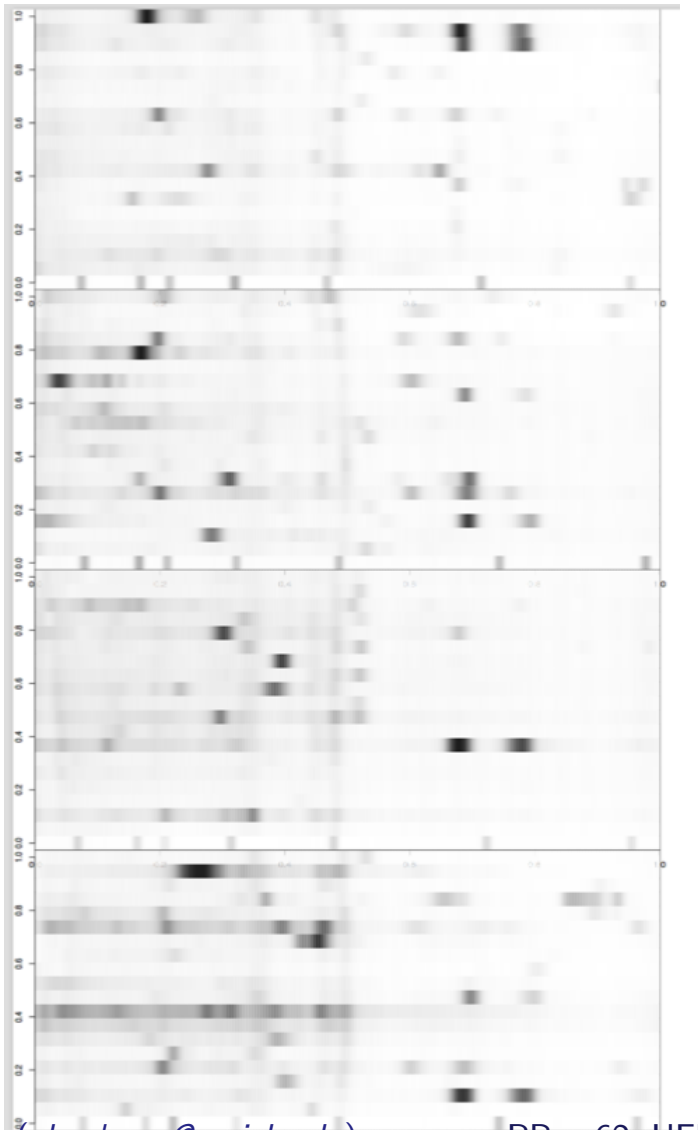
Step I-C: Aligned High-Frequency Intensity Data

Before



Step I-C: Aligned High-Frequency Intensity Data

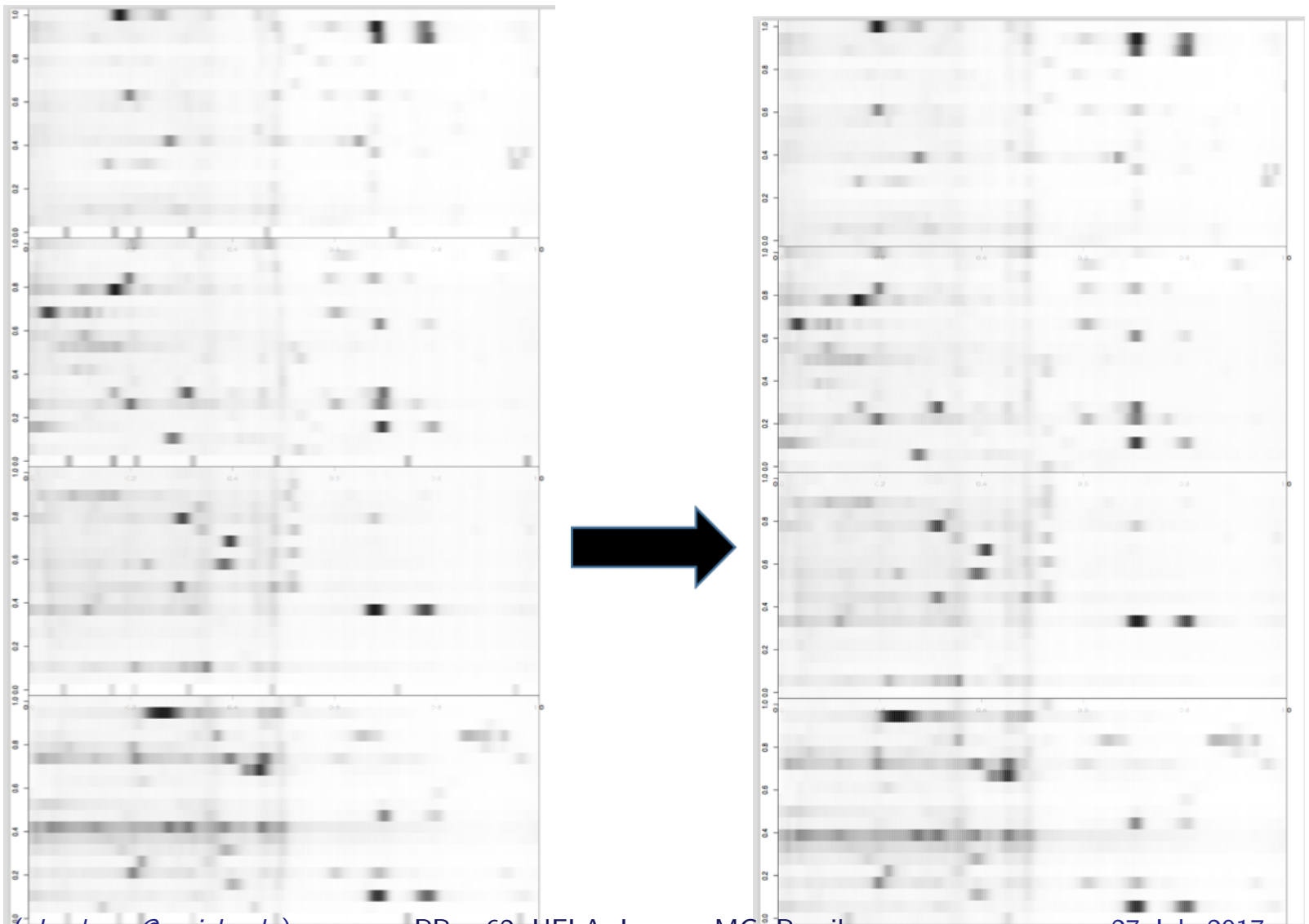
Before



Step I-C: Aligned High-Frequency Intensity Data

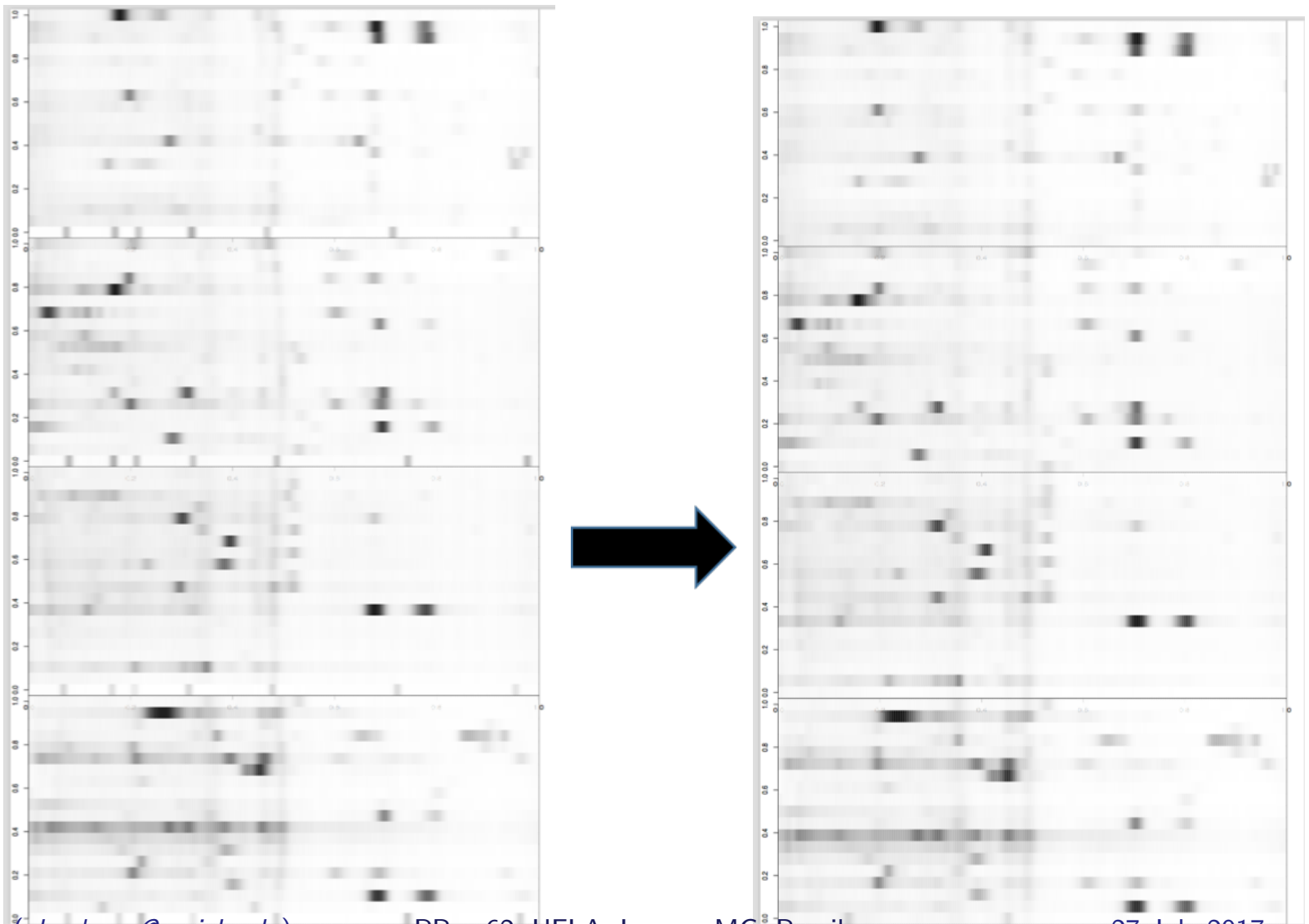
Before

After



Step I-C: Aligned High-Frequency Intensity Data

Before → note the curvatures are removed → After



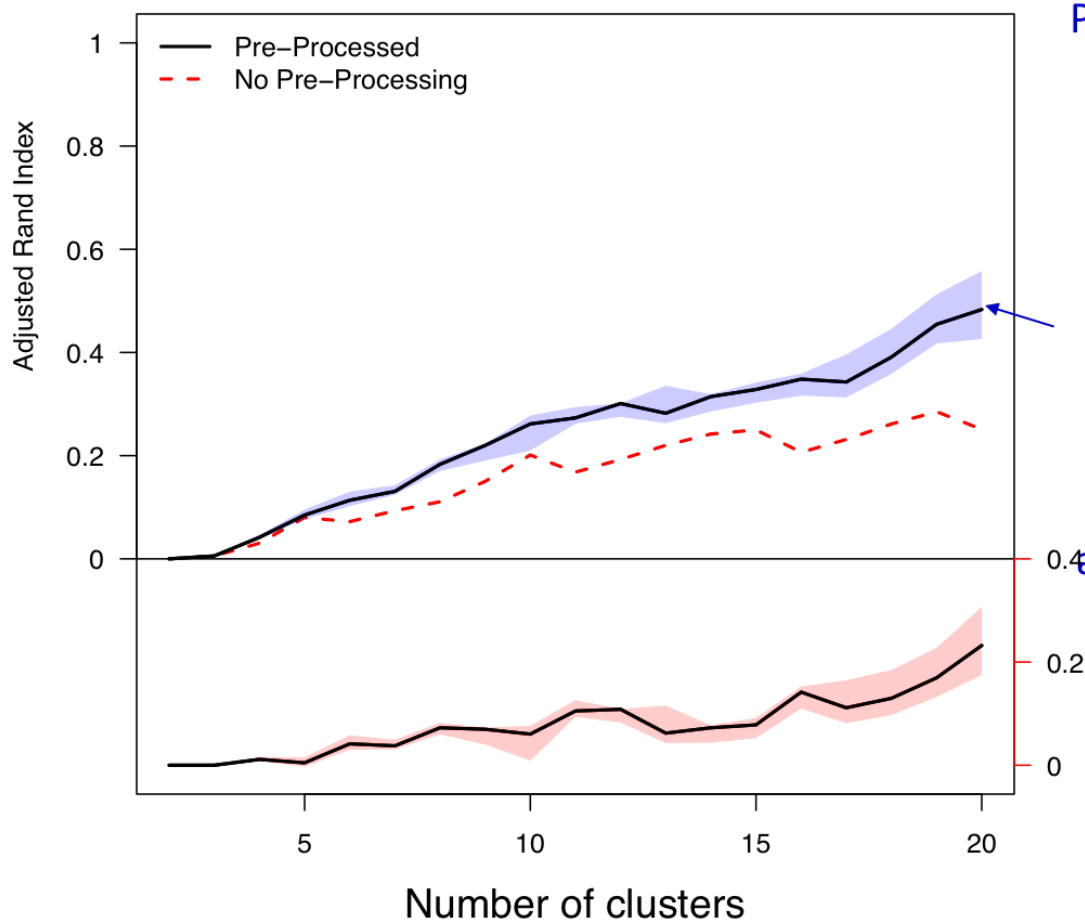
Data

Scleroderma

- Long-term clinical objective: find autoantibody signature that subsets autoimmune disease patients into groups with more homogeneous phenotypes and trajectories
- Sera from well-characterized patients with scleroderma and an associated cancer from Johns Hopkins Scleroderma Center database
- Data
  1. Known clustering: two replicate GEA experiments on 20 samples
  2. Unknown clustering: non-replicate GEA experiment on 80 samples
- Steps:
  1. Pre-processing
  2. Clustering (into 2, 3, . . .,N groups) based on the pre-processed high-frequency intensity data (hierarchical clustering here)
  3. Evaluate the separation of the obtained clusters and compare them to the truth (known in the replicate experiment)

Pre-processing Improves the Accuracy of Cluster Estimation

Data with technical replicates; 20 samples, long- and short- exposures



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

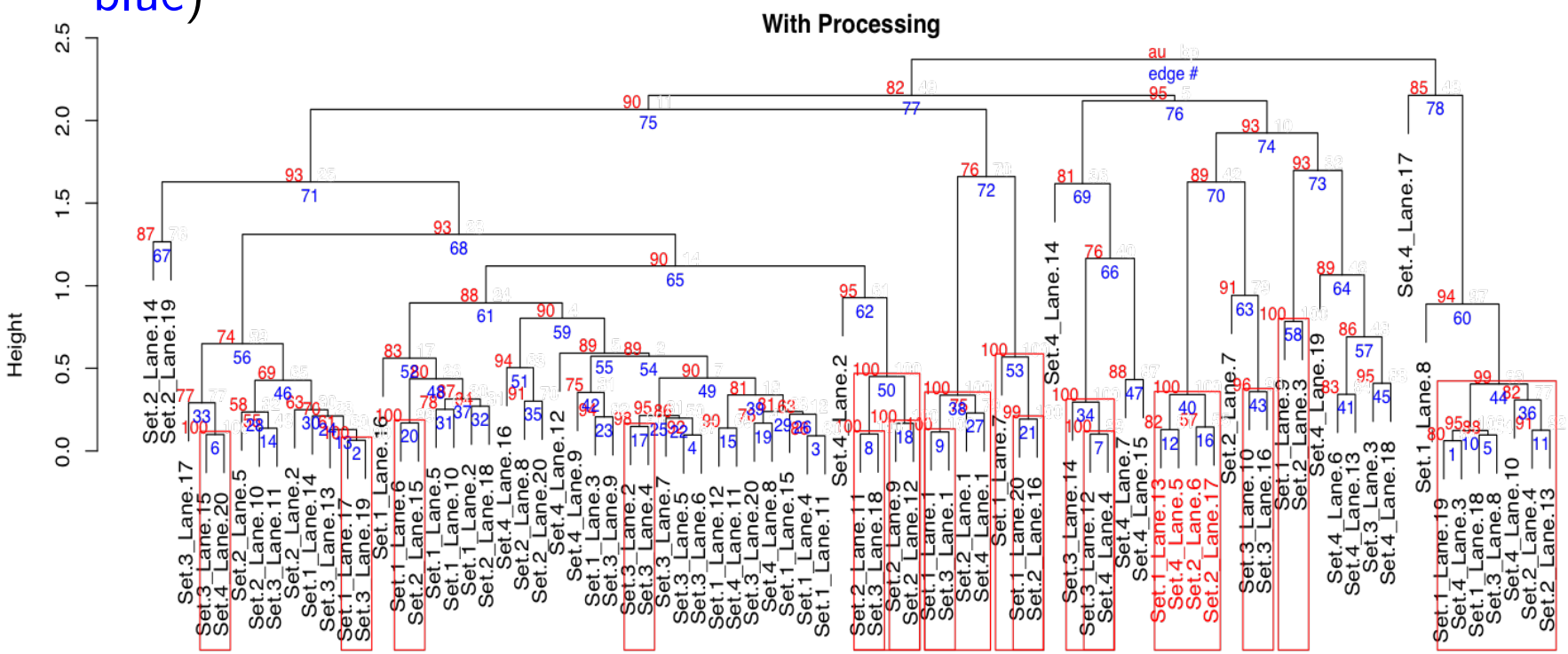
Recall that we obtain the posterior distribution of alignment $[\mathbf{Z} \mid P]$

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

Pre-processing Improves the Separation of Clusters

Data without Replicates; Hierarchical Clustering; Pre-processed vs Non-Pre-processed

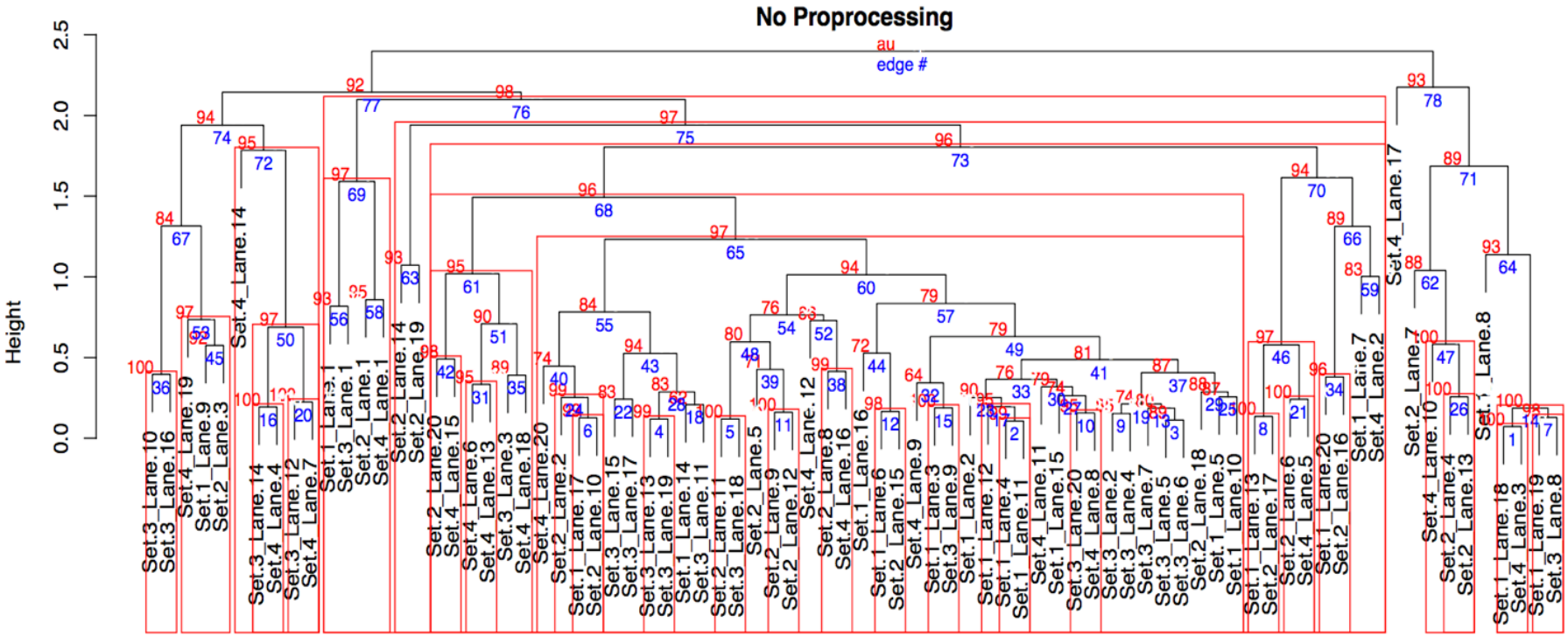
- Distance: Correlation-based distance; complete linkage
- Interpretation: adjacent terminal nodes in the tree $\rightarrow$ similar in AutoAntibody signatures
- Uncertainty: confidence levels by multiscale boostrapping (red numbers; ones > 95 are shown in red boxes; a numbering of the subtrees is shown in blue)



Pre-processing Improves the Separation of Clusters

Data without Replicates; Hierarchical Clustering; Pre-processed vs Non-Pre-processed

- Distance: Correlation-based distance; complete linkage
- Interpretation: adjacent terminal nodes in the tree $\rightarrow$ similar in AutoAntibody signatures
- Uncertainty: confidence levels by multiscale boostrapping (red numbers; ones > 95 are shown in red boxes; a numbering of the subtrees is shown in blue)



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
  1. Estimate autoantibody signatures
  2. The pre-processed data (Step I) can be the input of many subgroup discovery methods (Step II) including hierarchical clustering, latent class models and factor analyses
  3. Improves the accuracy of subgroup discovery
- Free publicly available 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; Based on the biology that autoantibodies recognize protein complexes.

Thank You!

Funding
Patient-Centered Outcome Research Institute [PCORI ME-1408-20318]
Hopkins Individualized Health Initiative

Some References (More at: `zhenkewu.com`)

- **Wu Z**, Casciola-Rosen L, Shah AA, Rosen A, Zeger SL (2017+).
  Estimating AutoAntibody Signatures to Detect Autoimmune Disease Patient Subsets.
  Minor Revision for *Biostatistics*. http://biorxiv.org/content/early/2017/04/18/128199.
- **Wu Z**, Deloria-Knoll M, Hammitt LL, and Zeger SL, for the PERCH Core Team (2015).
  Partially Latent Class Models (pLCM) for Case-Control Studies of Childhood Pneumonia Etiology.
  Journal of the Royal Statistical Society: Series C (Applied Statistics). 65:97-114.
- **Wu Z**, Deloria-Knoll M and Zeger SL (2016a).
  Nested Partially-Latent Class Models for Estimating Disease Etiology from Case-Control Data.
  Biostatistics, 18 (2): 200-213. doi:10.1093/biostatistics/kxw037.