PRECISESADS

imi innovative medicines initiative

efpia

Genetic clustering in autoimmune diseases

Robust models to enhance diagnoses, treatments, and drug developments

2017.06.06

Thomas Charlon

Prof. S Voloshynovskiy

Dr. J Wojcik

Quartz Bio

Outline

Problem: genetic autoimmunity challenges

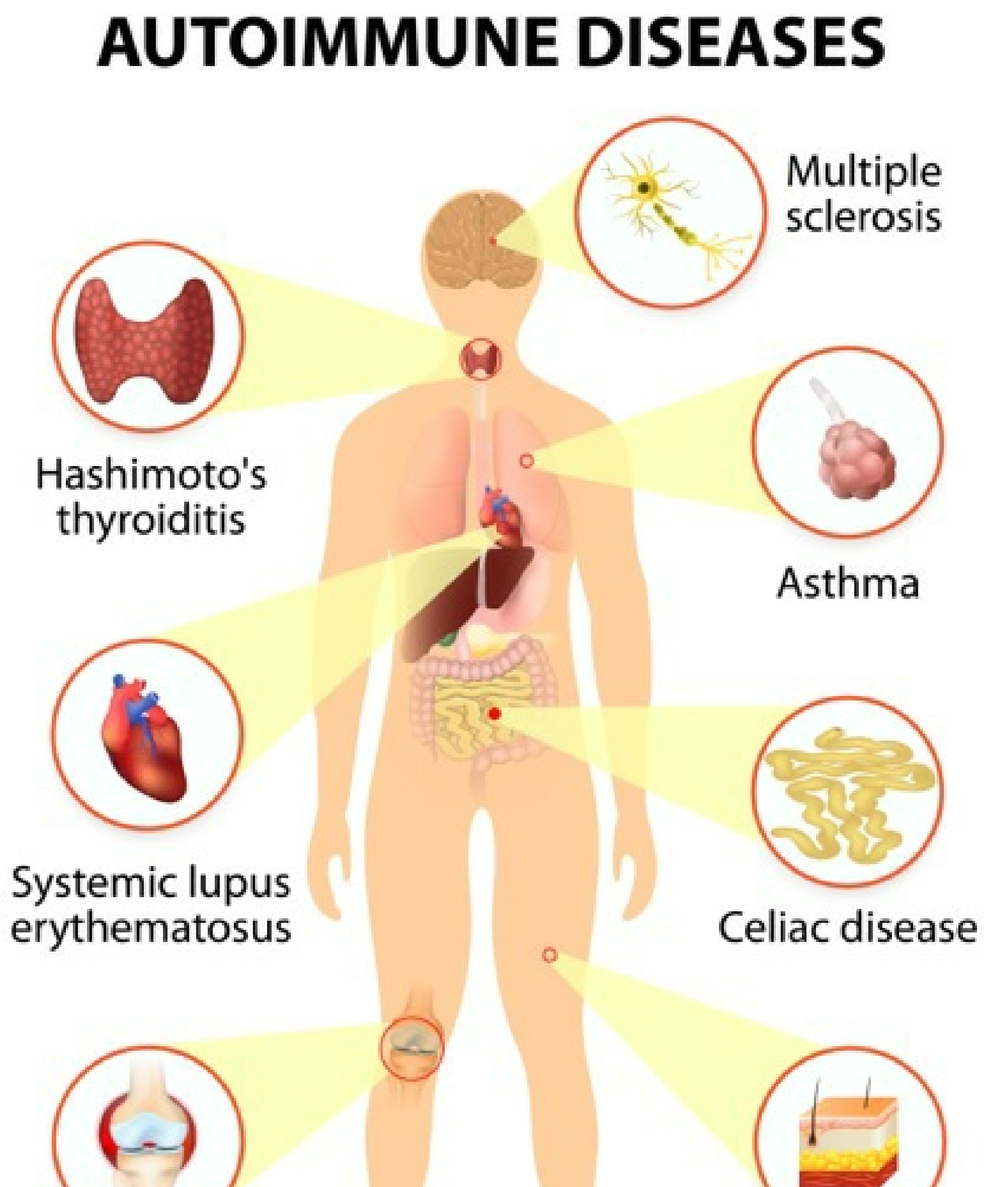
Hypotheses: clustering of associated markers and sparse modeling

Axis 1: Gaussian mixture model sub-sampling

Axis 2: Kernel projections and nearest neighbor models

Future works

Problem formulation



Systemic autoimmune diseases (SADs)

Severe chronic inflammations with variable symptoms and difficult diagnosis, general population affected at 1%.

5 types in PreciseSADs: systemic lupus erythematosus (SLE), systemic sclerosis (SSc), rheumatoid arthritis (RA), Sjögren's syndrome (SjS), and one group of mix and undifferentiated cases.

Challenges for clinicians

- Symptoms are shared accross diseases (e.g. kidney disease): difficult diagnosis
- In each disease, symptoms vary: treatments hard to develop

Goals of PreciseSADs project

Identify molecular signatures to enable clinicians to tailor therapies.

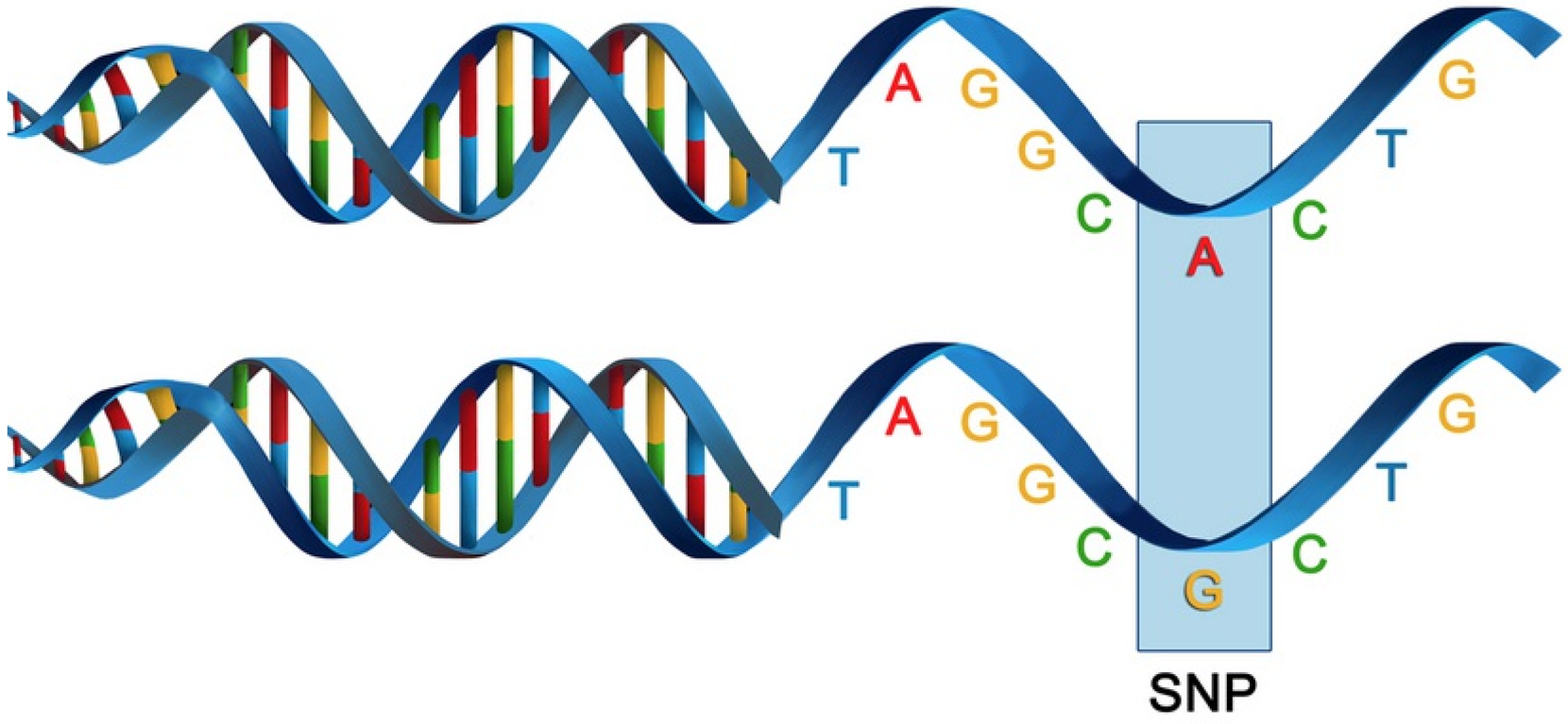
Reclassification of the patients and discovery of biomarkers

Genetics

Single Nucleotide Polymorphisms (SNPs)

Most common genetic variations, ~200,000 measured with PreciseSADs microarrays.

Ternary categorical values, often modeled as numerics 0, 1, and 2.



Genetics of systemic autoimmune diseases

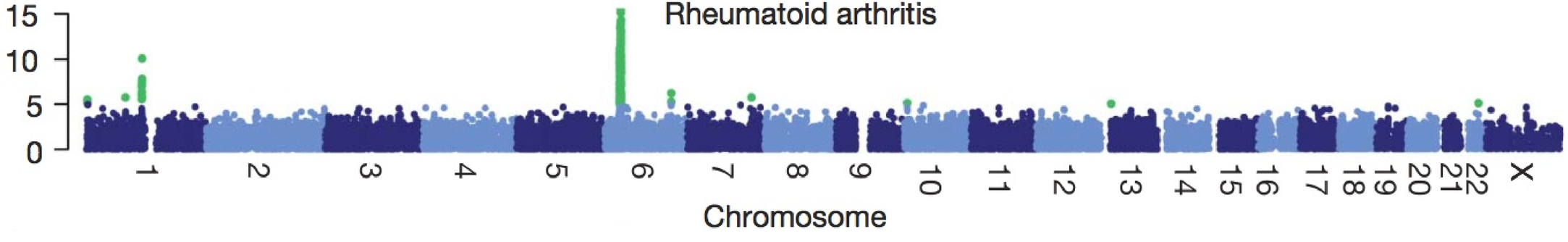
Patients' families and twins are 5-20 times more affected than general population.
Genetic component is estimated at 20-50%.

Association studies

Association studies compared markers frequencies between patients with one disease and healthy controls. Markers only explain ~2x more risk in SADs.

One large region, HLA on chromosome 6, contains the markers with highest risks.

Below, association study with one SAD on ~5,000 cases and controls. Green indicate associated markers. HLA is the most associated (WTCCC - 2007).



Previous work and reported results

Genome-wide PCA summarization

When considering all 200,000 markers, clustering associated to clinical centers is revealed.

We developed a method to filter out this population clustering, based on summarizing physically close contributors to principal components.

The method produces ~600 features and clustering is not associated to centers, but nor to diseases.

Clustering of risk markers

Gaussian mixture model (GMM) on 400 patients revealed diseases associated groups.

Now, 550 patients and additional HLA risk markers imputed by CSIC (Granada, Spain).

Hypotheses

Unsupervised clustering of the risk markers

- 80 markers of the most associated region are used for clustering patients
- Optimal number of clusters with BIC metric and sub-sampling for robustness

Kernel and sparse models to increase clustering

- Increase clustering metrics and robustness
- Experiment with more markers and other associated regions

Research axis 1: Gaussian mixture model on risk markers

80 markers of HLA are measured for 550 patients of ~5 diseases

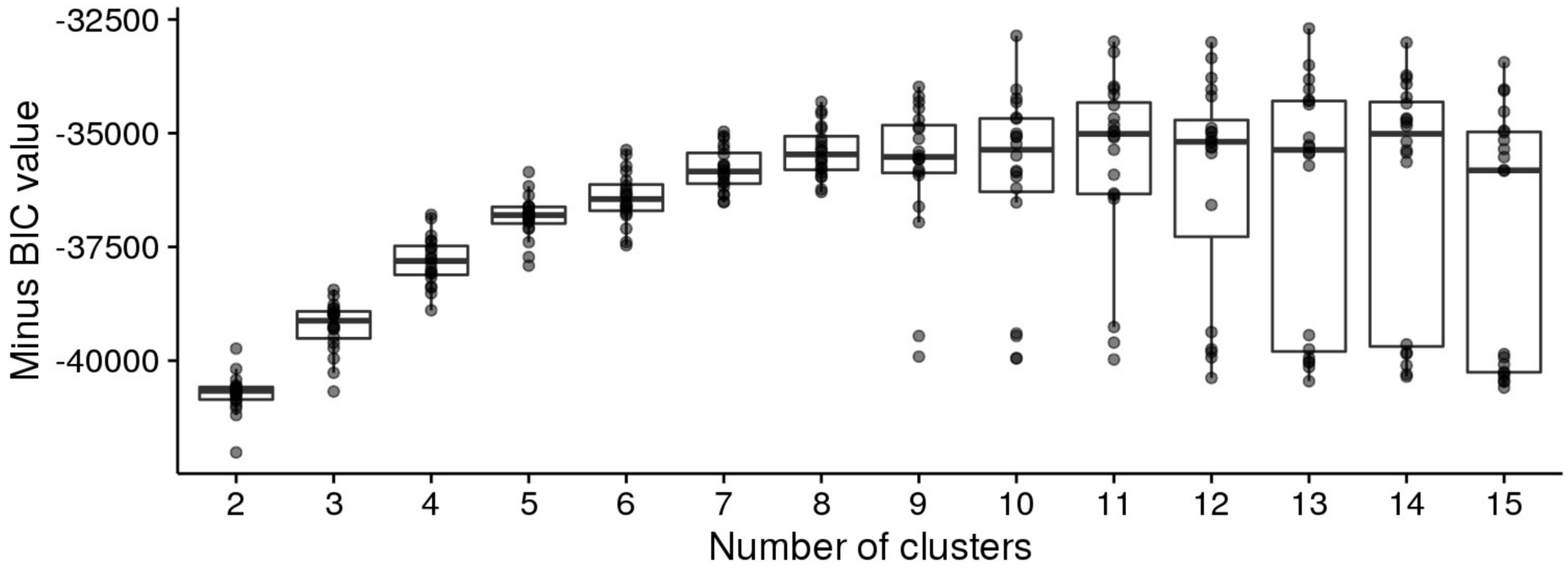
GMM clustering and sub-sampling

Goal: Determine number of clusters

Method: highest minimun of sub-sampled BIC, goodness of fit metric

- computed 20x on 90% of patients for 2-15 clusters

Result: optimal number is 8



Clusters investigation: enriched diseases

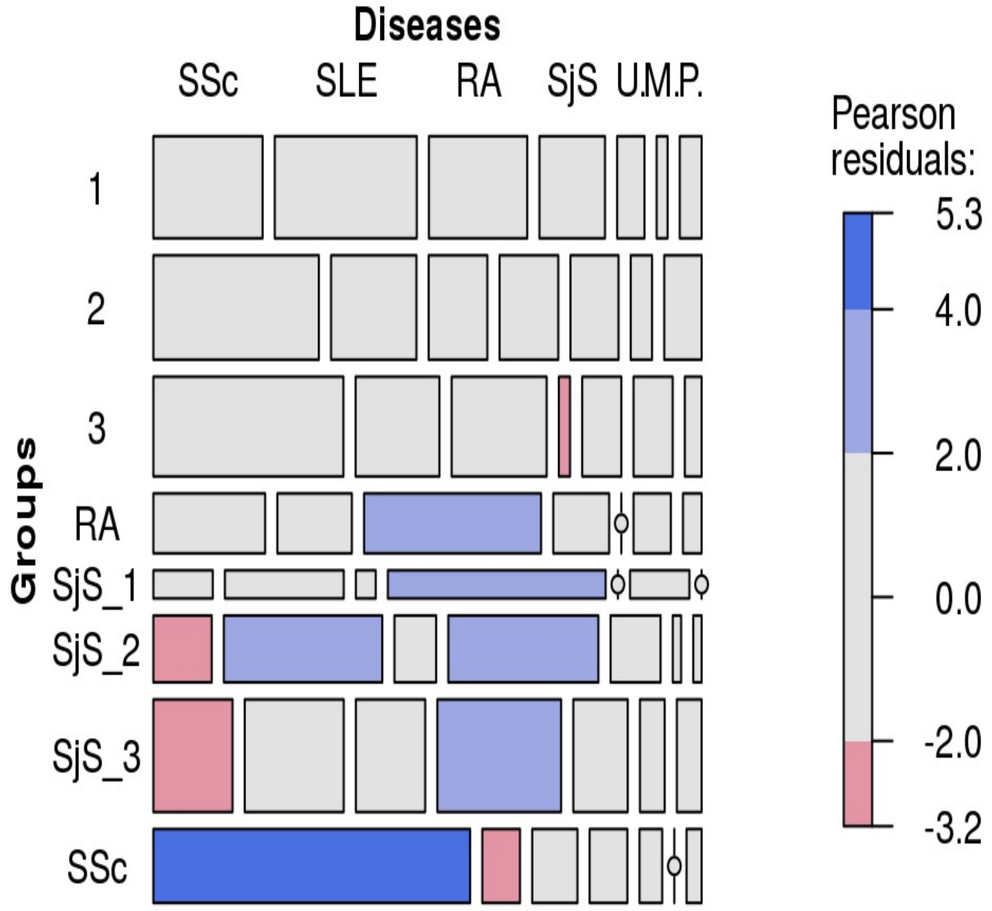
Goal: identify groups with enriched diseases

Method: Pearson residuals

$$\forall \text{ group } i,\ \text{disease } j,$$
$$r_{i,j} = \frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}},\ E_{i,j} = \frac{n_i n_j}{n}$$

Result: SSc strongly enriched, SjS and SLE in one group, SjS in 2 others, RA in 1

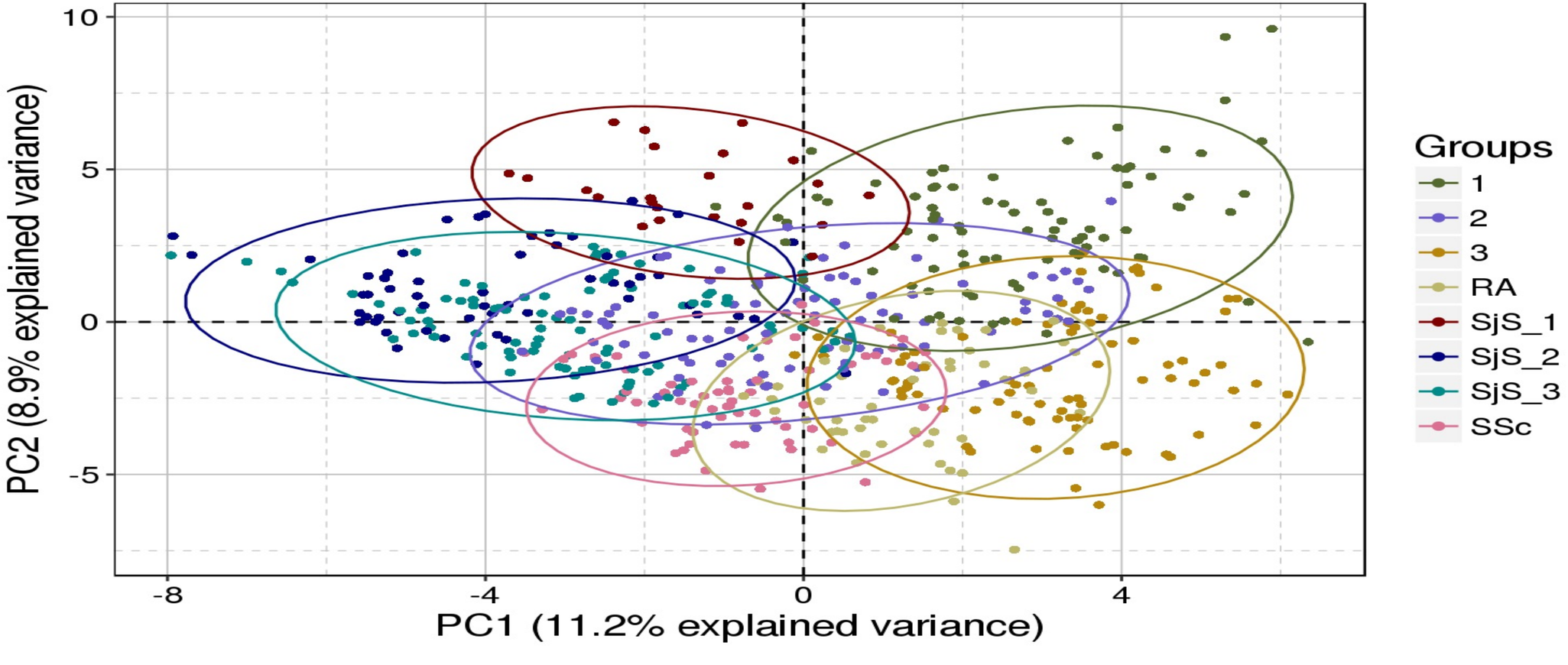
- cell height: number of patients in group.
- width: number of patients with disease relative to the group.
- color: enriched above 2 (blue), depleted below -2 (red)



Clusters investigation: PCA

To investigate the groups' similarities, they are mapped to the global principal component analysis (PCA).

SjS groups are together in top left, RA-SSc in bottom, groups 1 and 3 right



Discriminating markers: regressions

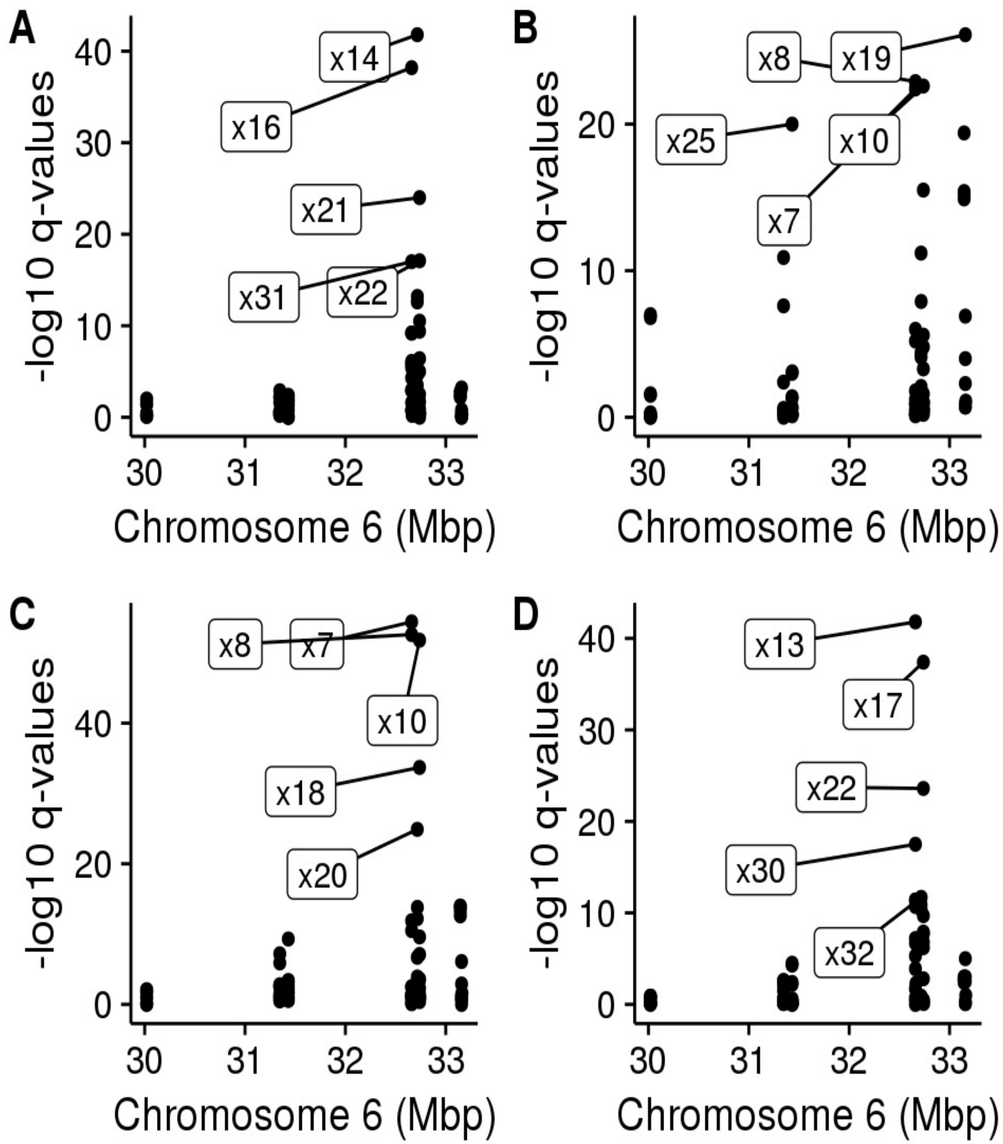
Goal: identify markers associated with groups

Method: regressions between markers and each group against all others

- usual association study method

Figure:

- x: marker position, y: associations
- A: RA, B: SjS_2, C: SjS_3, D: SSc

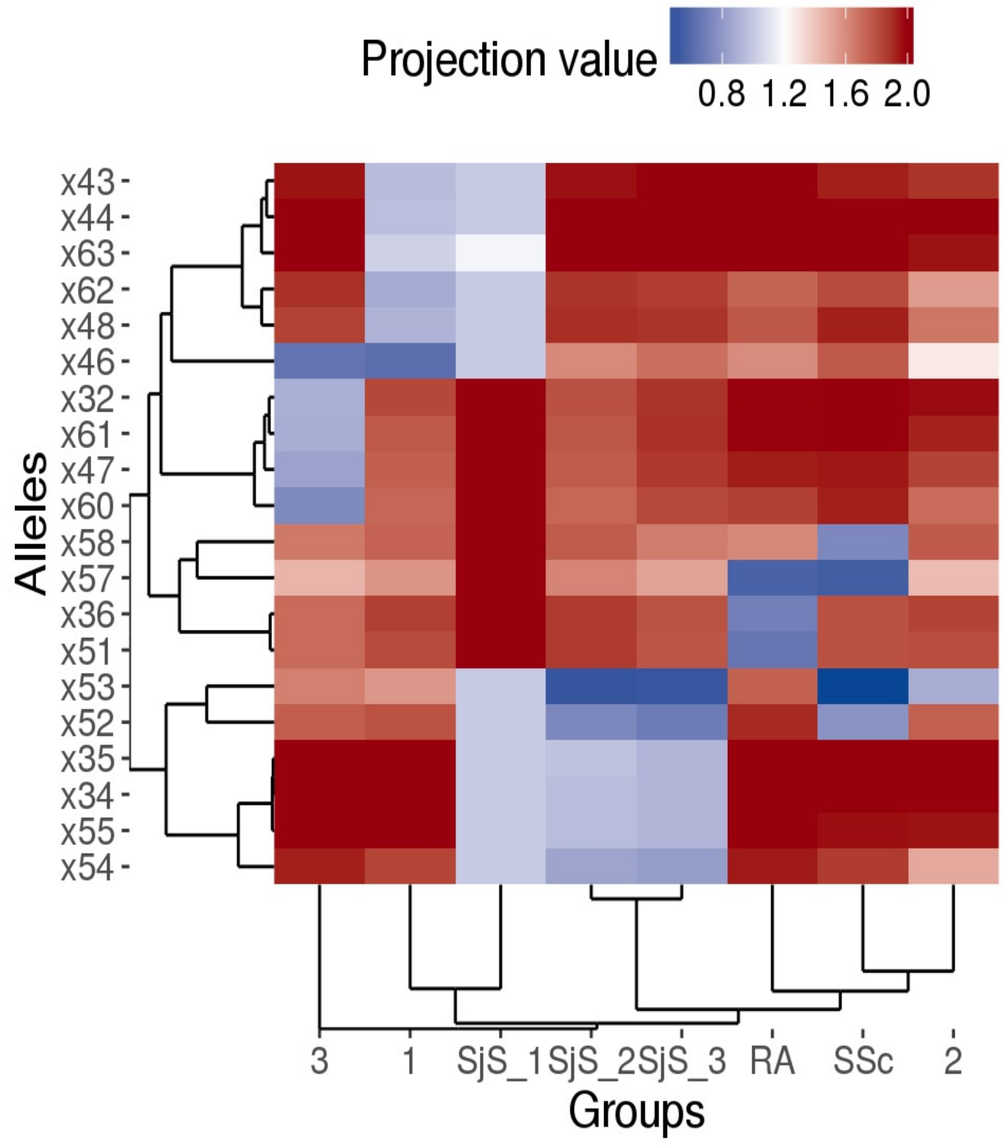


Discriminating markers: heatmap

Goal: identify correlations in discriminating markers

Method: GMM projection matrix heatmap

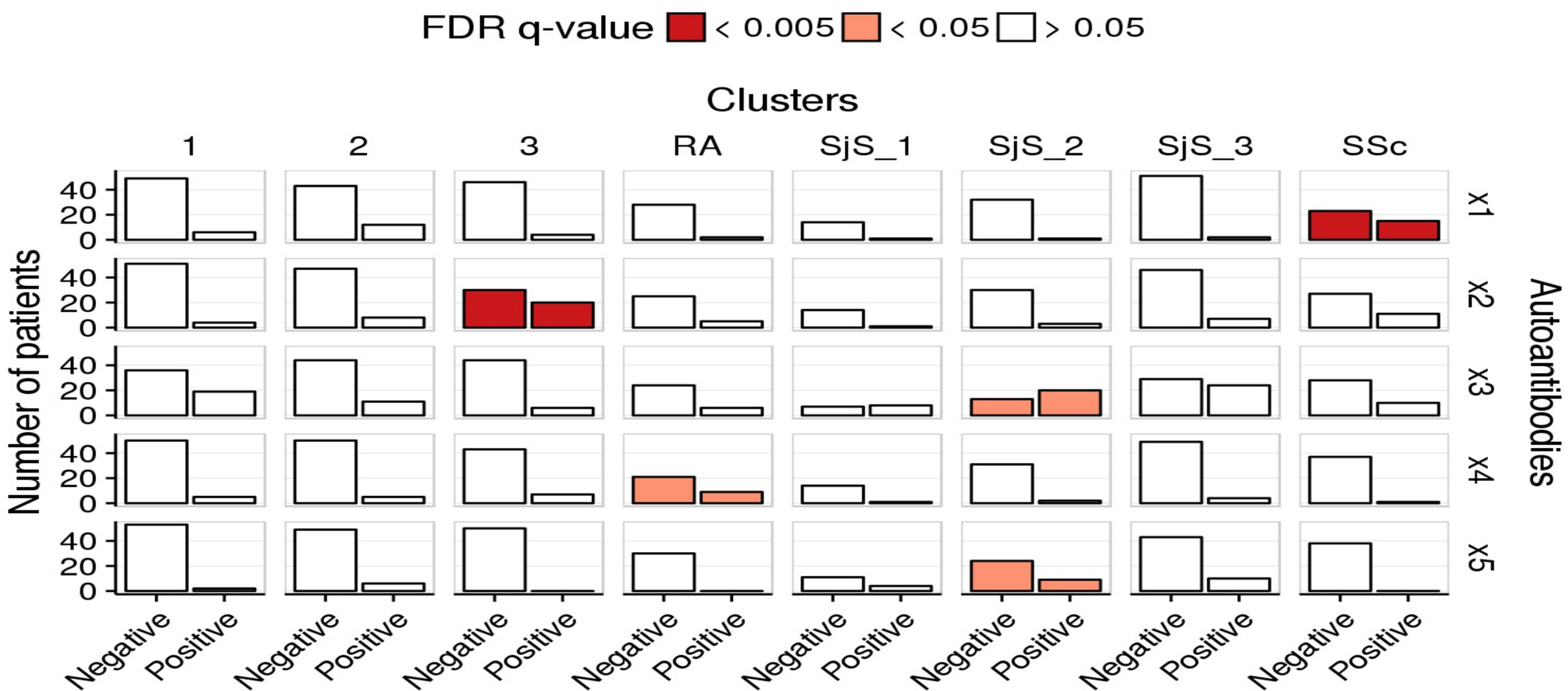
- subset to 20 markers with most variance of projection between groups
- ordered by hierarchical clustering



Autoantibodies associations

Goal: identify autoantibodies and proteins presence enrichment in groups

Result: associated with SSc, with SjS, and with RA as expected. Last is a SSc related autoantibody.



Supplementary analyses

Include controls in clustering and compare compositions of clusters:

- 9 similar clusters including 1 associated with controls

Associations of other risk SNPs with groups:

- only HLA SNPs are associated

Diseases not enriched in any groups:

- discriminating markers identified, UCTD similarities with SjS, MCTD with RA

Conclusions

Novel profiles with specific autoantibodies are identified in genetic markers

Submission to Arthritis and Rheumatology (Wiley), internal review

Research axis 2: kernel and sparse models

Increase robustness of clusters compositions when adding or removing patients

Discover clusters in larger number of markers

Methods

Kernel projections: polynomial, gaussian, laplace

$$(\alpha X^T X + c)^{degree},\ \exp -\frac{|X - X^T|^2}{2\sigma^2},\ \exp -\frac{|X - X^T|}{\sigma}$$

Sparse encoding, with euclidean distance or inner product:

$$X^d = X^T X$$

- 1 Hard: k nearest neighbors set to 1, others set to 0

$$\forall i, j,\ X^h_{i,j} = \begin{cases} 1 & \text{if } X^d_{i,j} \text{ in } i\ k\text{-}neighbors \\ 0 & \text{else} \end{cases}$$

- 2 Soft: exponent transform of neighbors distances, others 0

$$\widetilde{X_{i,j}} = \exp \frac{X^d_{i,j}}{\sigma},\ X^{\mathrm{s}}_{i,j} = \frac{\widetilde{X_{i,j}}}{\sum_{j=1}^{n} \widetilde{X_{i,j}}}$$

- 3 Epsilon: exponent transform of distanes and threshold by mean of all distances

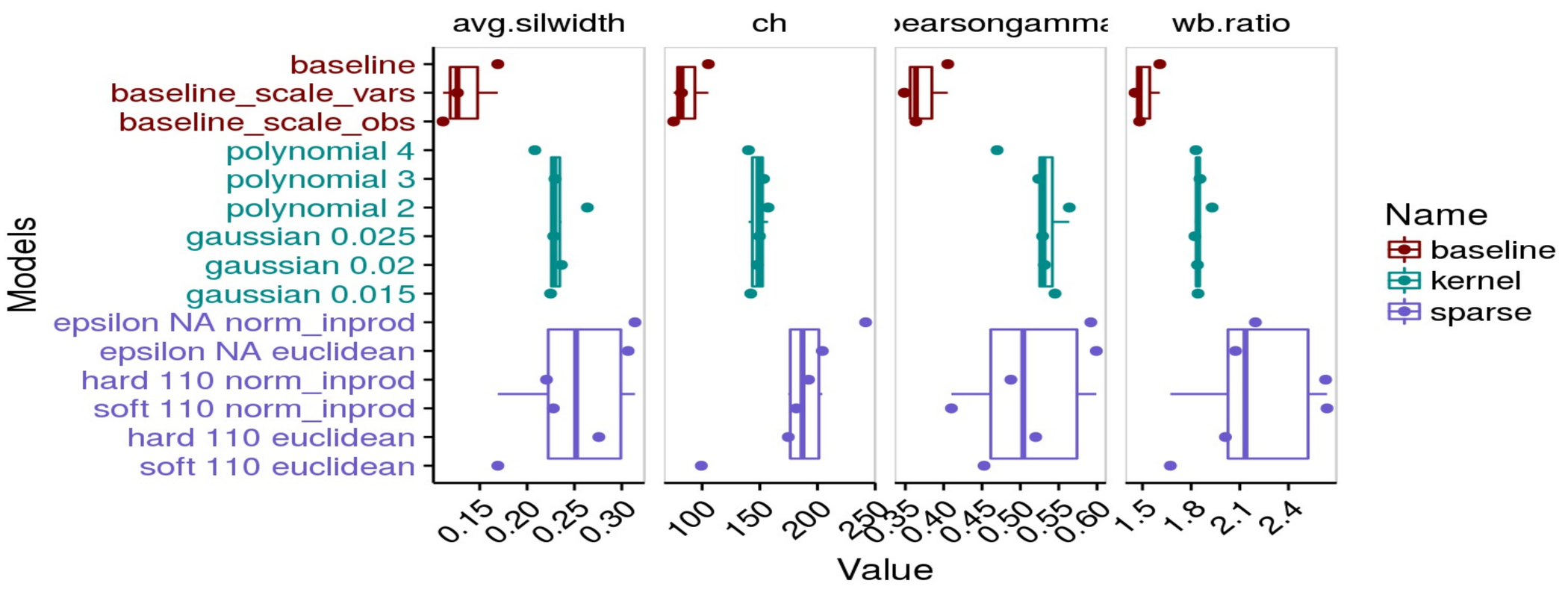
$$\forall i, j,\ X^e_{i,j} = max(mean(X^{\mathrm{s}}_{i,j}), X^{\mathrm{s}}_{i,j})$$

Distance metrics

Method: metrics from distances in 5 principal components

- within-between ratio: ratio of distances inter and intra clusters
- average silhouette width, ch, pearsongamma

Result: sparse models outperform kernels and baseline



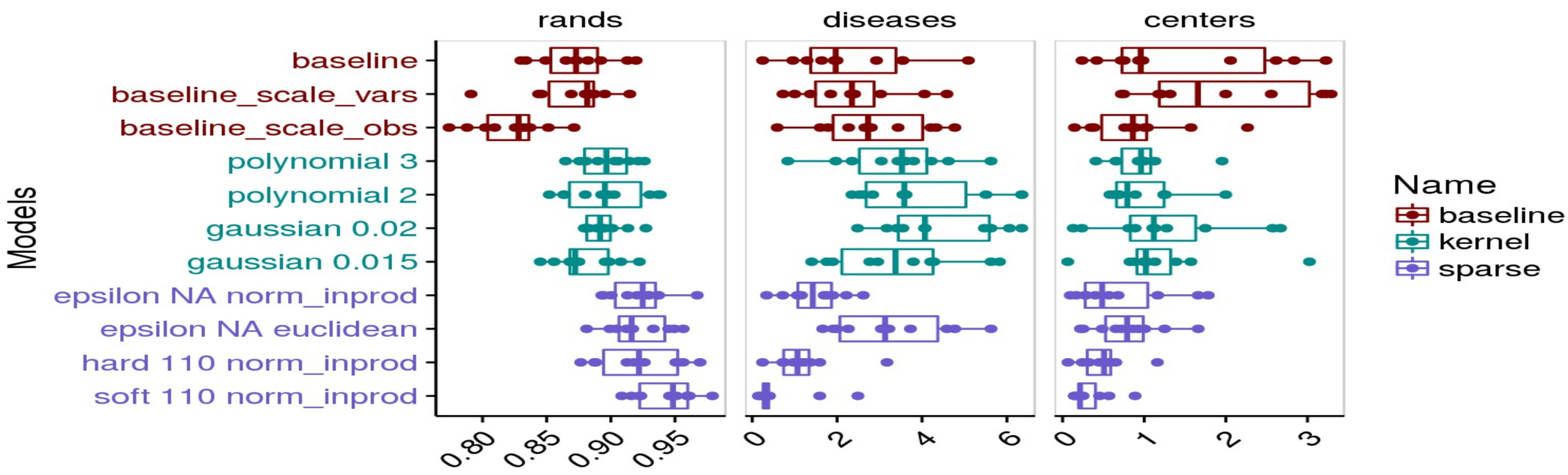
Robustness and associations

Goal: compare robustness and associations with diseases/centers

Method: sub-sampling with 0.5 ratio

- robustness: similarity with clustering of all patients (rand)
- associations: sum of pearson residuals (chi2)

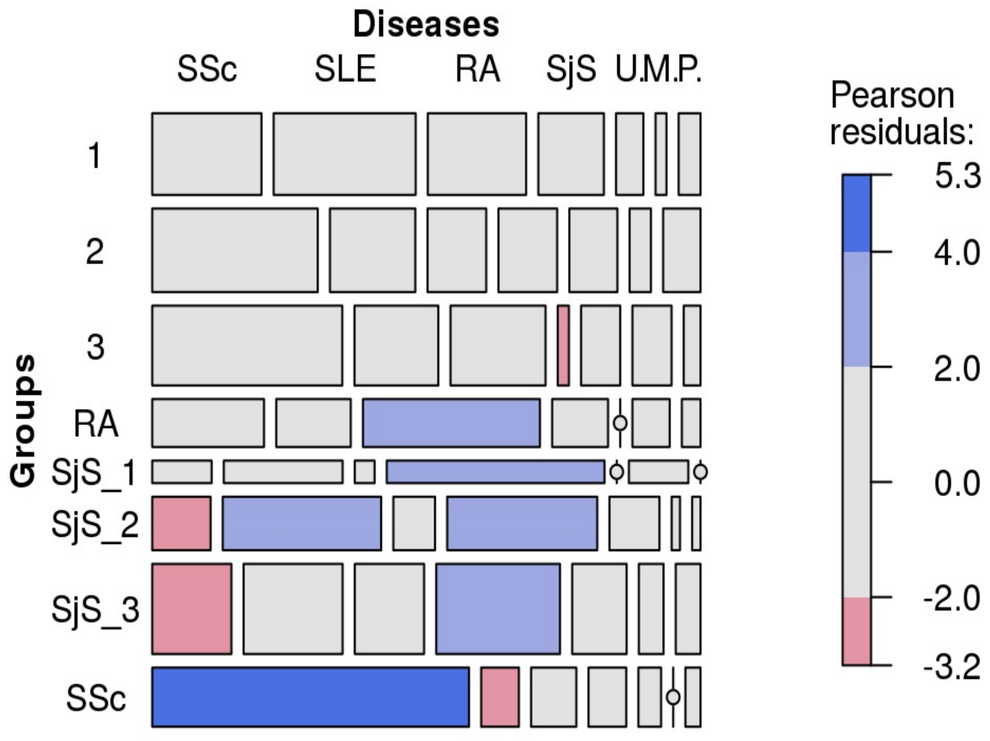
Result: sparse models more robust and euclidean epsilon more associated with diseases and less with centers than baselines.

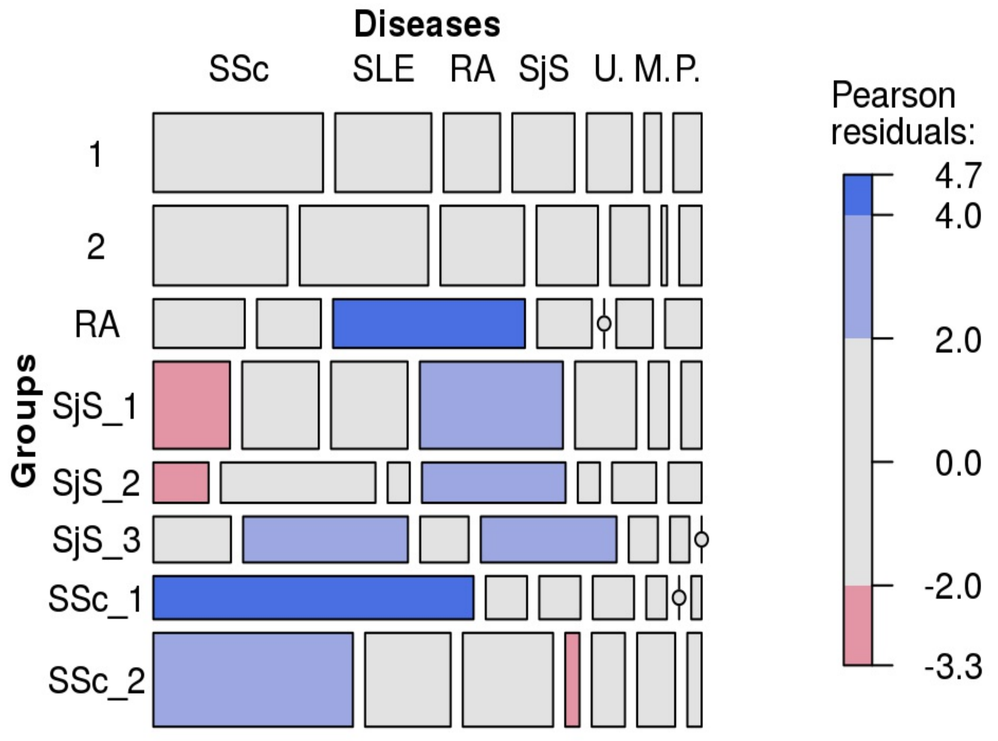


Clusters investigation

Compared to baseline (left), in the transformed domain (right)

- RA was more associated
- Second SSc cluster was revealed

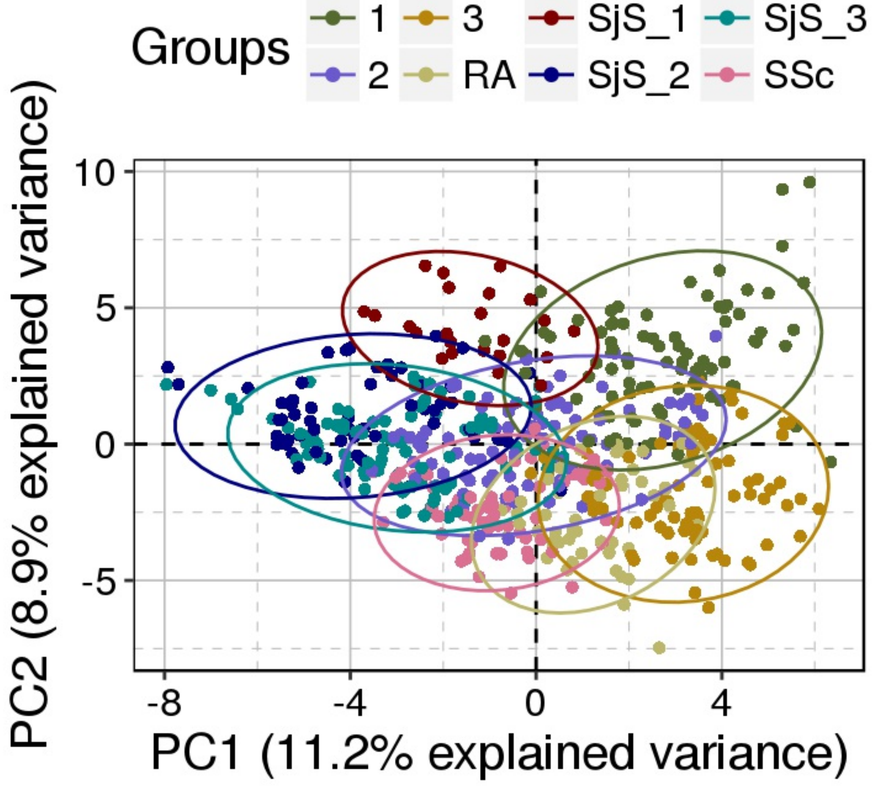


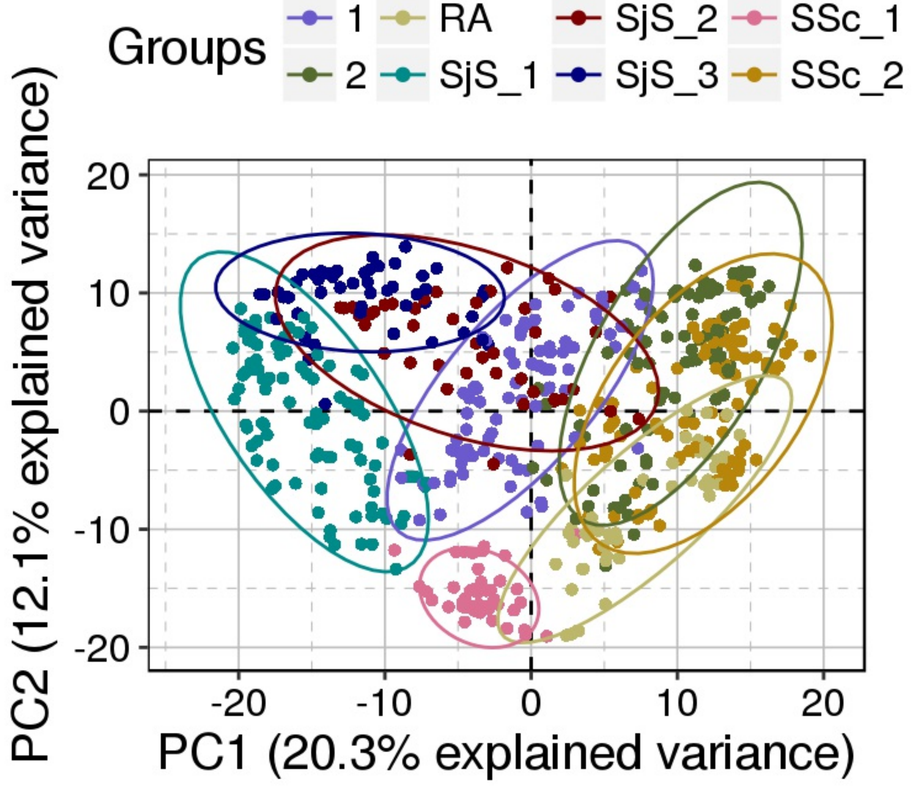


Clusters investigation

Compared to baseline (left), in the transformed domain (right)

- Groups in PCA were more compact

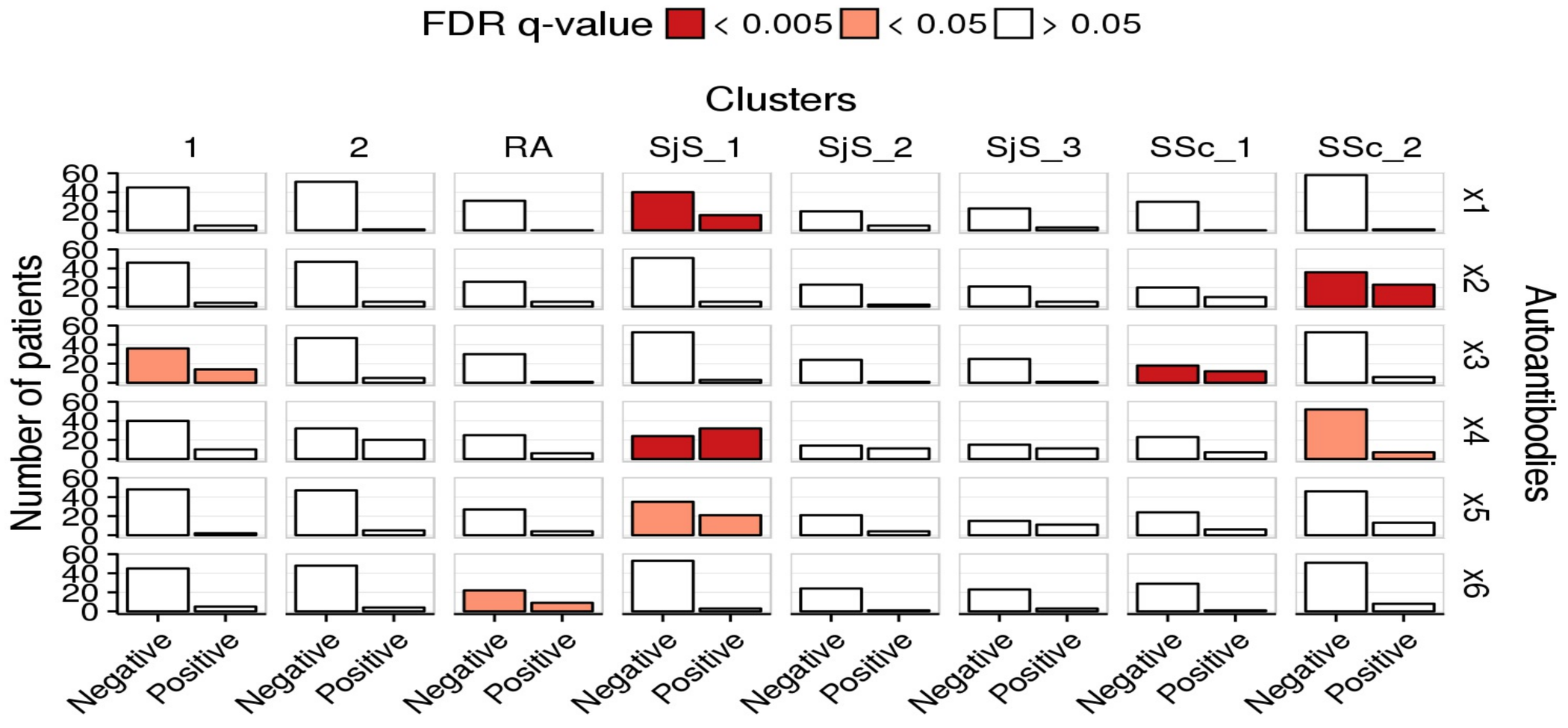




Autoantibodies associations

Autoantibodies are also more enriched:

- Novel association, 2 more strong associations



Supplementary datasets

Kernels and sparse models perform slightly better than baseline but no interesting results

Conclusions

Clustering and robustness are increased, novel cluster enriched in SSc, novel autoantibody associated

Dissemination in preparation

Conclusions

Summary

Gaussian mixture model clustering of the risk markers

Novel profiles associated with specific autoantibodies revealed in genetic markers

Medical article in review

Kernel and sparse models to increase clustering

Clustering and robustness are increased, novel cluster enriched in SSc, novel autoantibody associated

Dissemination in preparation

Current works

Further investigation of GMM and sparse models

More patients are being collected and results may evolve.

Inclusion of other associated regions and SNPs.

Integrative clustering with other biomarkers

Cell counts, methylation, and other biomarkers are also measured in the patients.

Methylation clustering results revealed similarities with HLA results.

Thank you for your attention