Lecture 7
Clustering + Scalability

DSCI 532, Data Visualization II

January 23, 2019

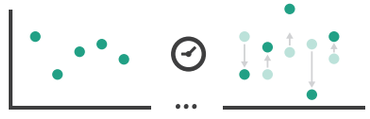
Cydney Nielsen

Senior Designer, Microsoft
Adjunct Professor, UBC Department of Computer Science

Dealing with complexity

Manipulate

- Change

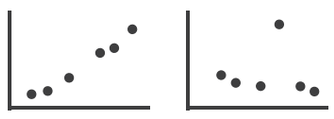


- Select

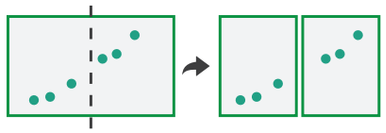
- Navigate

Facet

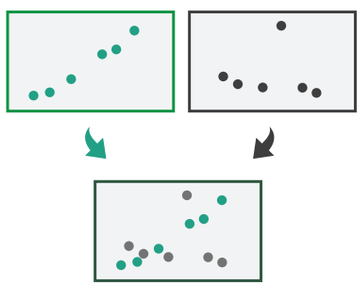
- Juxtapose



- Partition



- Superimpose



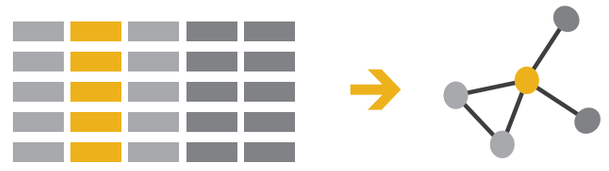
Reduce

- Filter

- Aggregate

- Embed

→ *Derive*



Overview

- **Clustering**
  - A way to visualize structure in a large datasets
- **Scalability**
  - **Perceptual scalability** - What to do when you have more data points than pixels
  - **Interactive scalability** - How do you make your visualization responsive

Clustering

Clustering

- As a dataset become large, you can no longer view all items
- Creating groups based on item similarity is an intuitive and highly useful approach to understanding your data
- Grouping items becomes non-trivial when the number of attributes is large (high-dimensional data)
- Clustering algorithms are powerful for identifying groups in high-dimensional data

Clustering tabular data

Attributes / Dimensions

Items

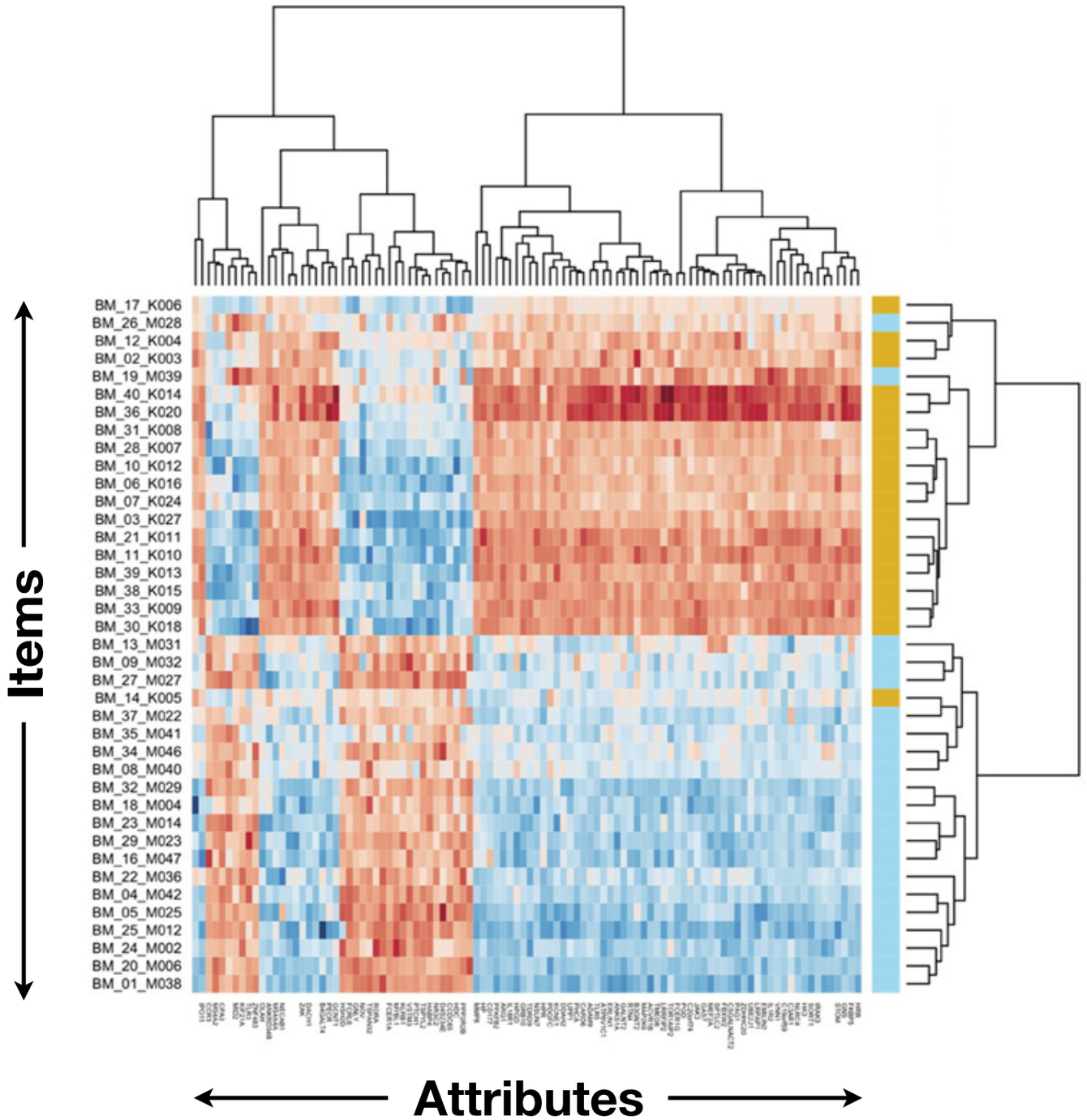
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.3	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
5.4	3.7	1.5	0.2	*I. setosa*
4.8	3.4	1.6	0.2	*I. setosa*
4.8	3.0	1.4	0.1	*I. setosa*
4.3	3.0	1.1	0.1	*I. setosa*
5.8	4.0	1.2	0.2	*I. setosa*

https://en.wikipedia.org/wiki/Iris_flower_data_set

Distance-based approaches

- Apply some metric to calculate the distance between two items (distance metric)
- Example - Hierarchical clustering is a common method
  - Bottom-up (agglomerative) clustering - each observation starts in its own cluster, and pairs of clusters are iteratively merged
  - Top-down (divisive) - all observations start in one cluster, and splits are performed recursively moving down the hierarchy

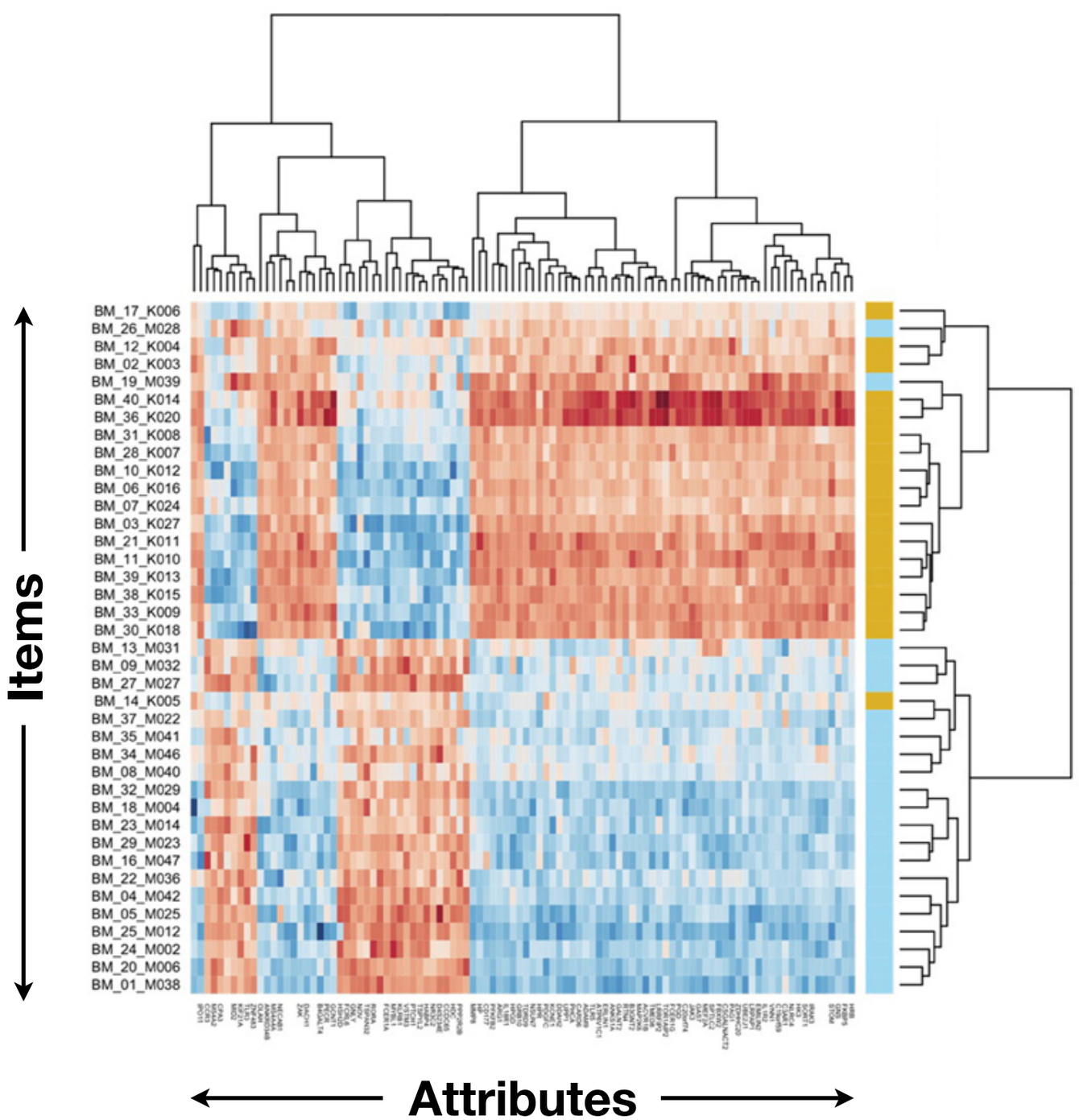
Heatmap + Dendrogram | Gene Expression Example



- Heatmap
  - In this example, colour represents the Z-score (number of standard deviations above or below mean for row; blue=high, red=low)
- Dendrogram
  - Dendrogram (tree diagram) shows merges
- Order matters
  - Cluster both rows (items) and attributes (columns) to reveal patterns

https://www.frontiersin.org/articles/10.3389/fnins.2013.00033/full

Heatmap + Dendrogram | Gene Expression Example



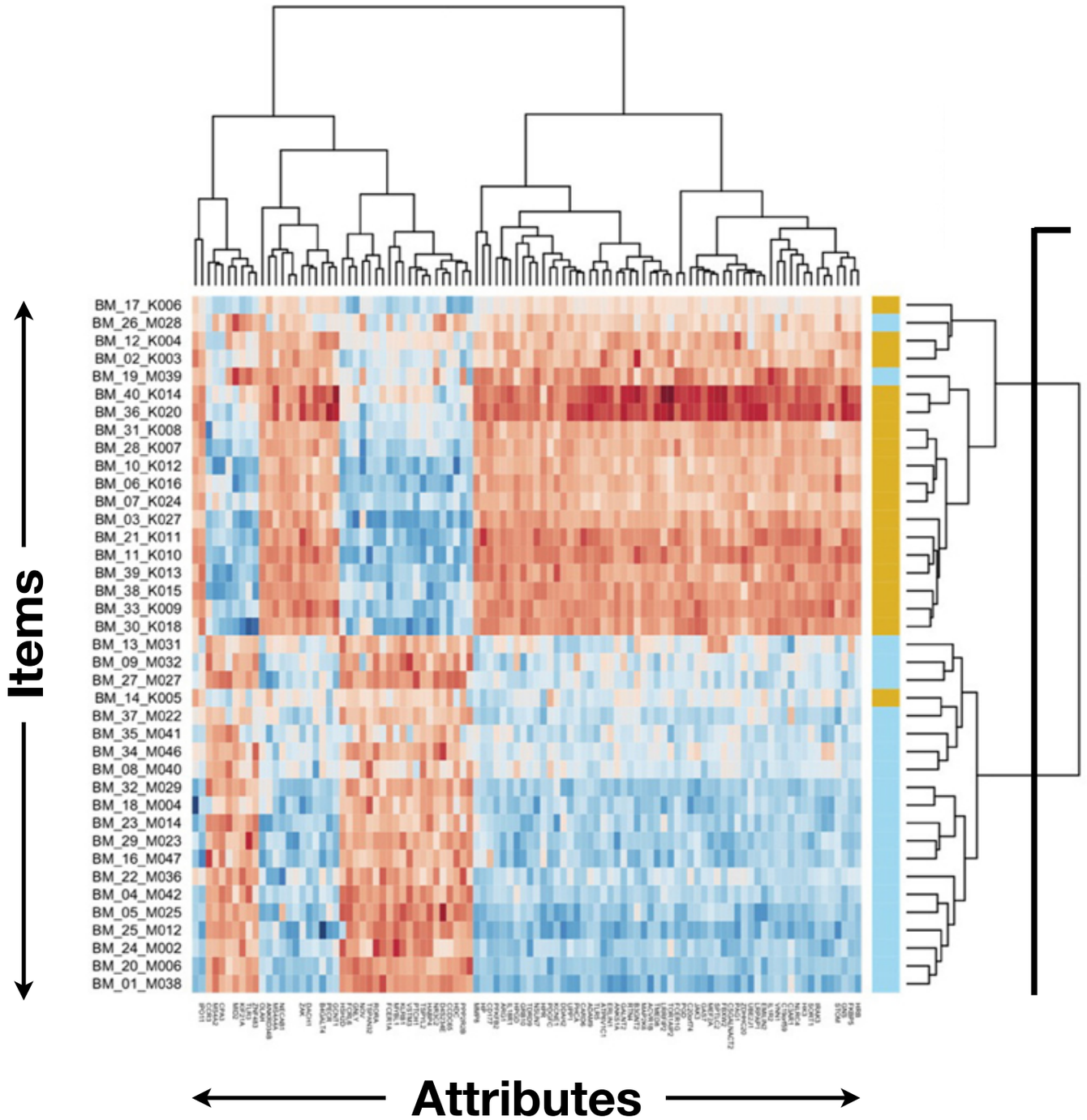
Meningitis patients

Healthy controls

Annotate rows with category labels (not used in the clustering) to see how they correlate with the clusters

https://www.frontiersin.org/articles/10.3389/fnins.2013.00033/full

Slicing a dendrogram

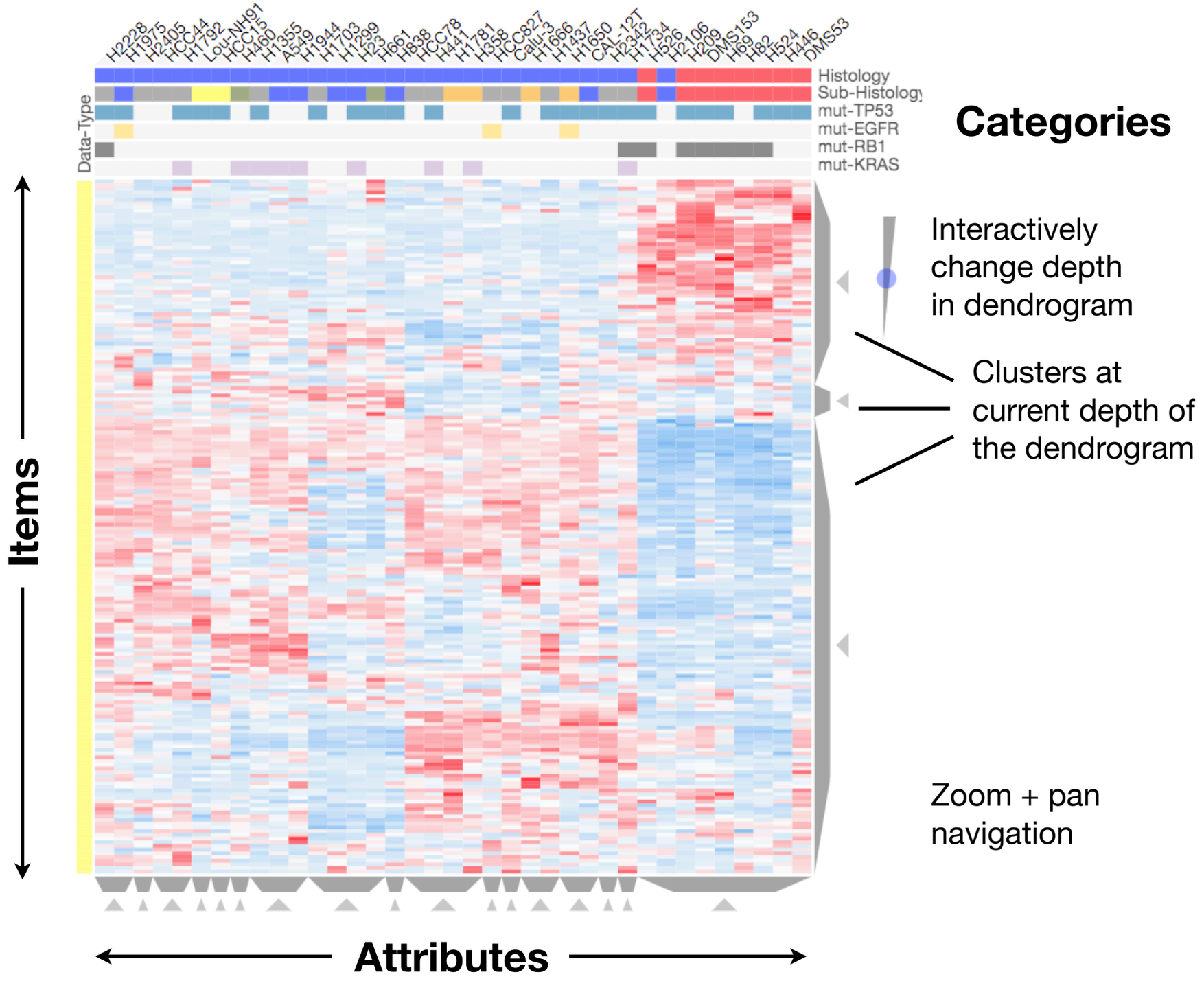


How many clusters?

Can cut at different depths of the dendrogram to get different numbers of clusters

https://www.frontiersin.org/articles/10.3389/fnins.2013.00033/full

Slicing a dendrogram | An interactive example



https://nbviewer.jupyter.org/github/MaayanLab/CST_Lung_Cancer_Viz/blob/master/notebooks/CST_Data_Viz.ipynb

Clustering sequences

Sequences of web site navigation (msnbc.com)

User	Sequence				
1	frontpage	news	travel	travel	
2	news	news	news	news	news
3	frontpage	news	frontpage	news	frontpage
4	news	news			
5	frontpage	news	news	travel	travel
6	news	weather	weather	weather	weather
7	news	health	health	business	business
8	frontpage	sports	sports	sports	weather
9	weather				

- **Variable length**
- **Heterogeneous**
  - Users tend to have vastly different browsing patterns
- **Large dataset size**
  - Server logs from msnbc.com for a 24 hr period produce roughly 1 million sequences

Visualization of navigation patterns on a Web site using model-based clustering

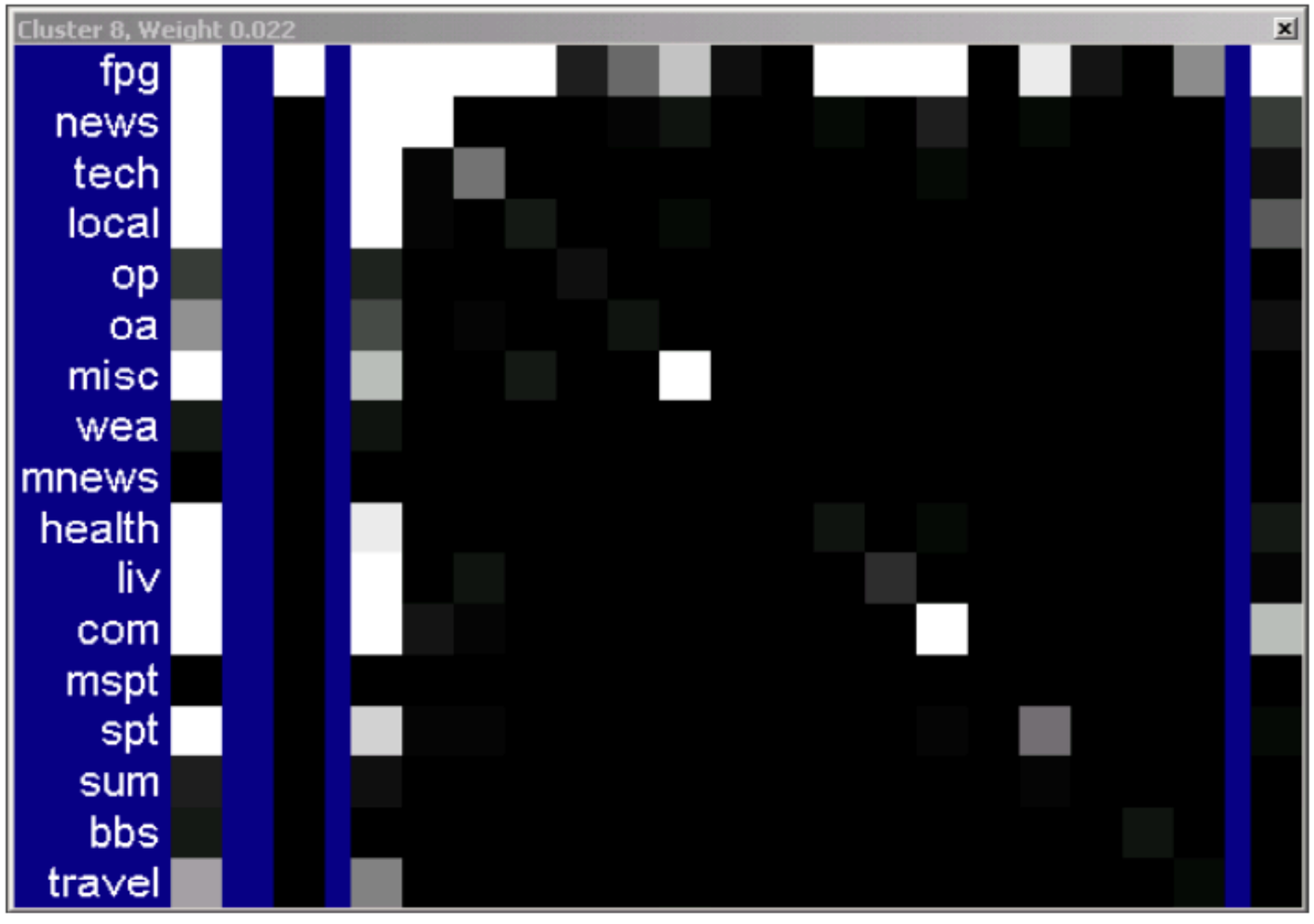
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.5622&rep=rep1&type=pdf

Clustering sequences | Model-based approach

- **Web navigation example - Markov model**
  - Model data as having been generated as follows:
    - A user arrives at the Web site and is assigned to a particular cluster with some probability
    - The behaviour of that user is then generated from a Markov model with parameters specific to that cluster
  - Use Expectation-Maximization (EM) to learn the proportion of users assigned to each cluster as well as the parameters of each Markov model
  - In doing so, assign each user to a cluster or fractionally to the set of clusters

Visualization of navigation patterns on a Web site using model-based clustering
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.5622&rep=rep1&type=pdf

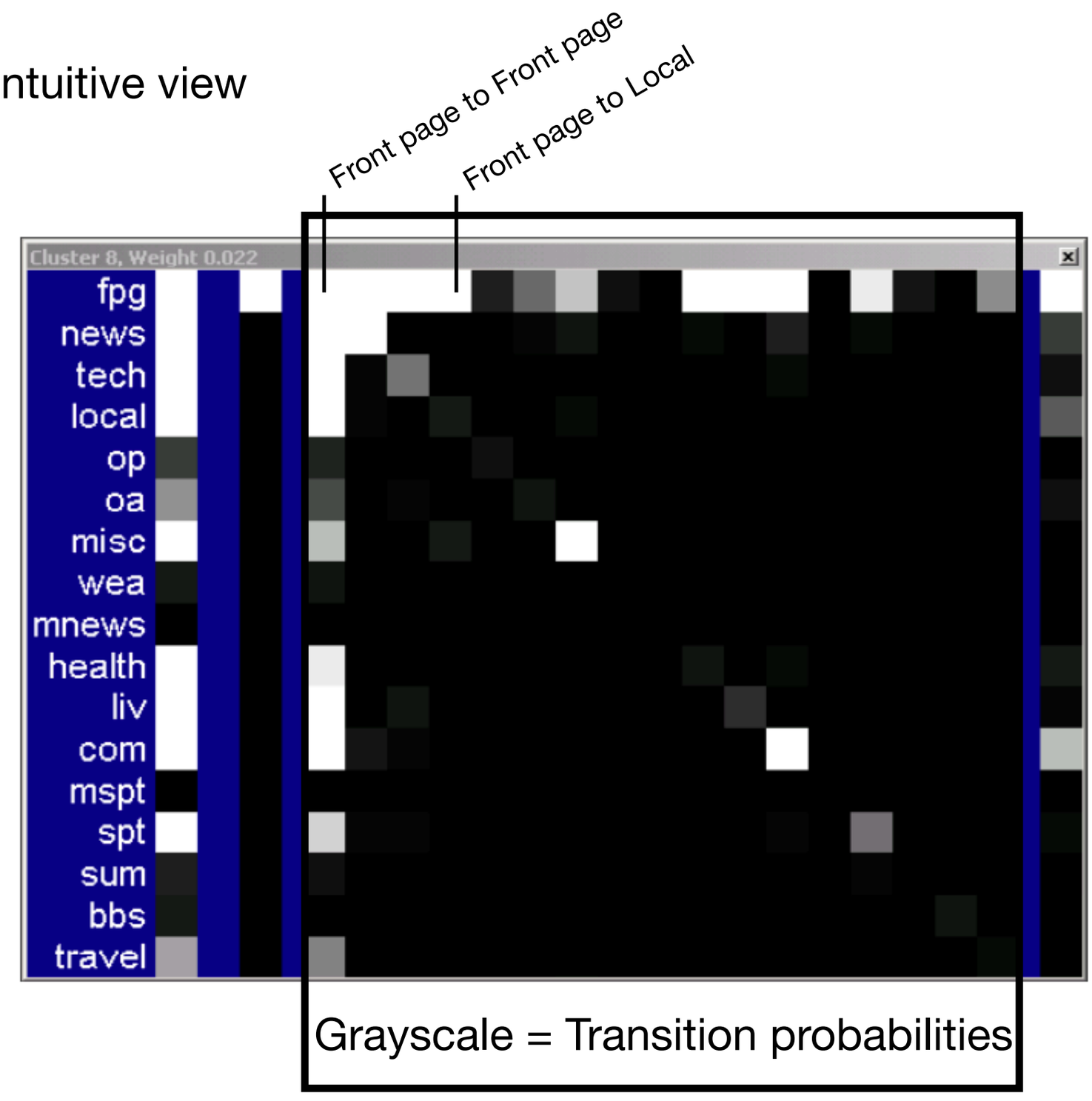
Visualizing the Markov model for one cluster



http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.5622&rep=rep1&type=pdf

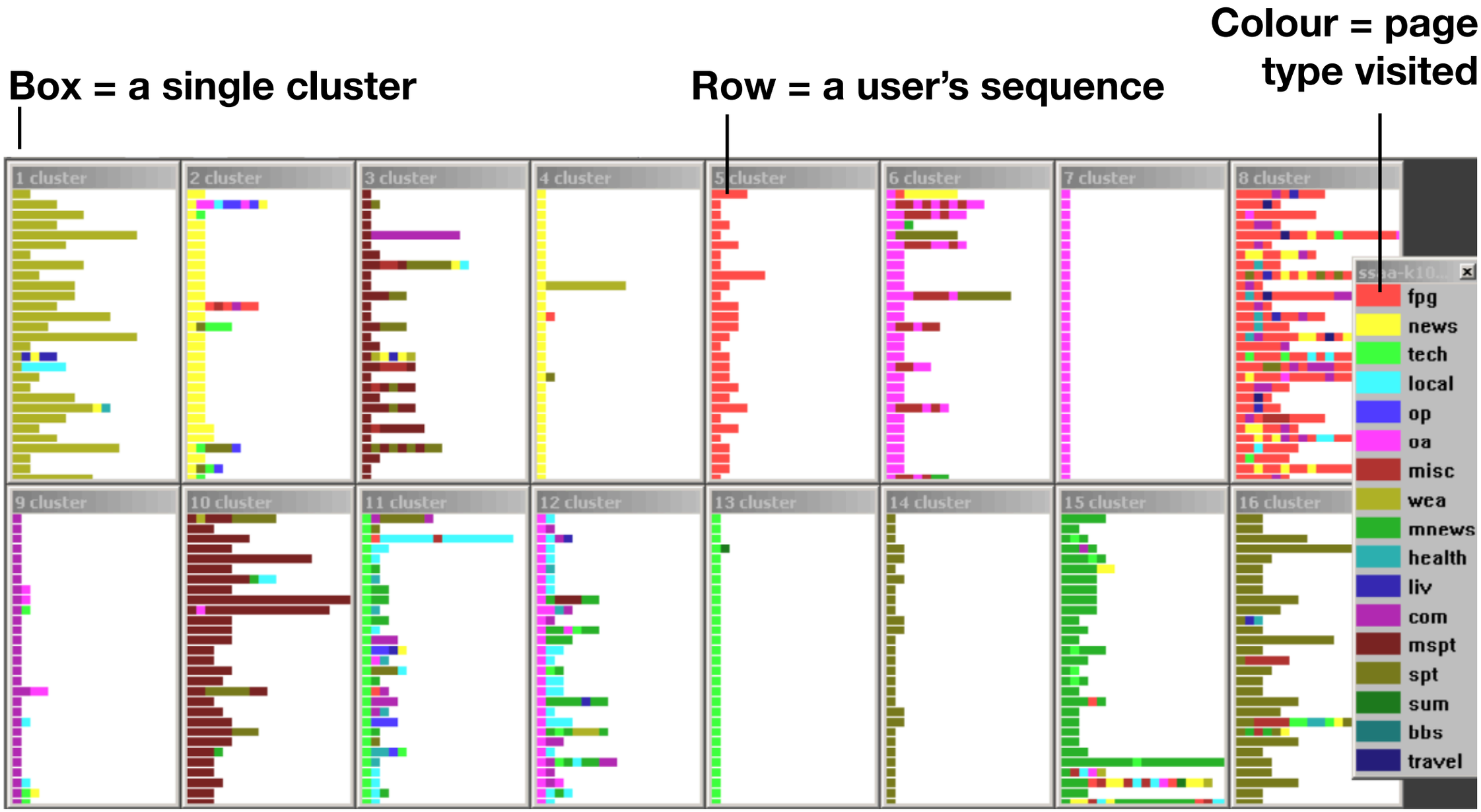
Visualizing the Markov model for one cluster

Not a very intuitive view



http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.5622&rep=rep1&type=pdf

Visualizing clusters + user sequences



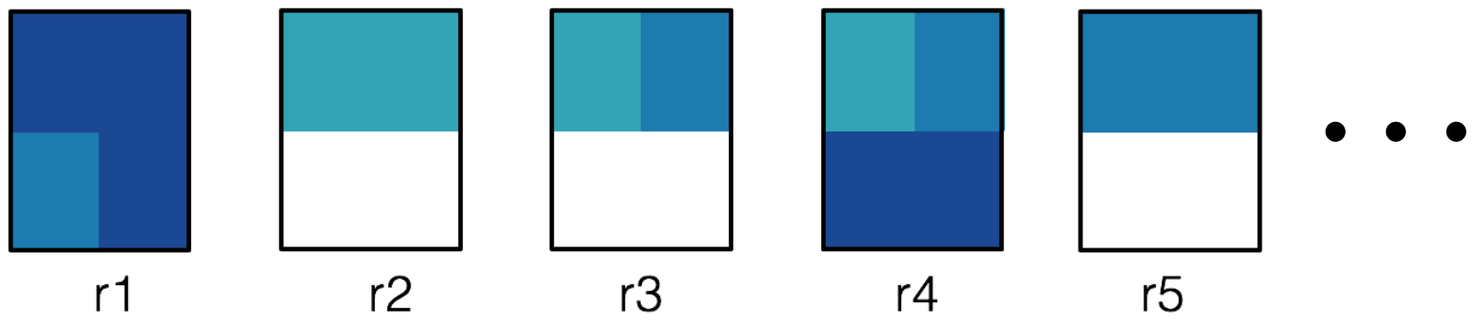
Can see overall patterns in each cluster

Intuitive to look at actual user sequences

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.5622&rep=rep1&type=pdf

Clustering | for navigation + exploration

A genomics example



Collect matrices of data from genomes

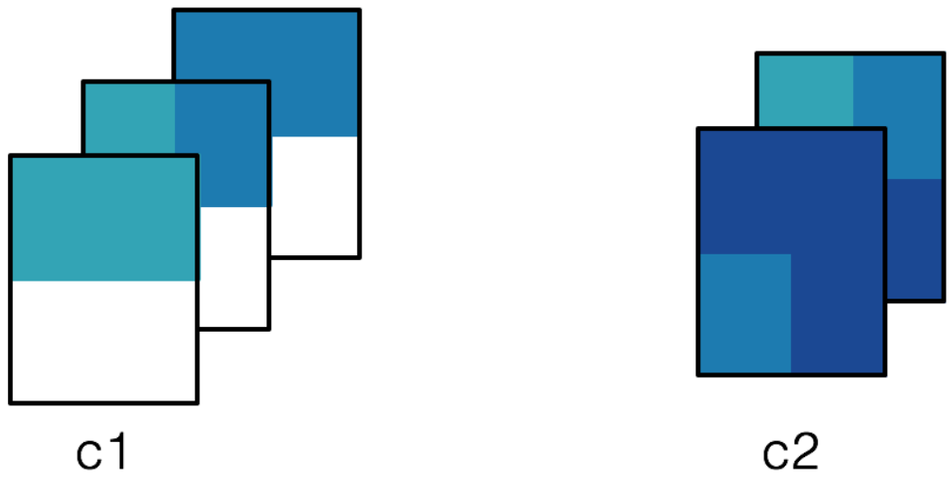
Tens to hundreds of thousands of these matrices

(dark blue = high values, white = zero)

http://sparkinsight.org/

Clustering | for navigation + exploration

A genomics example



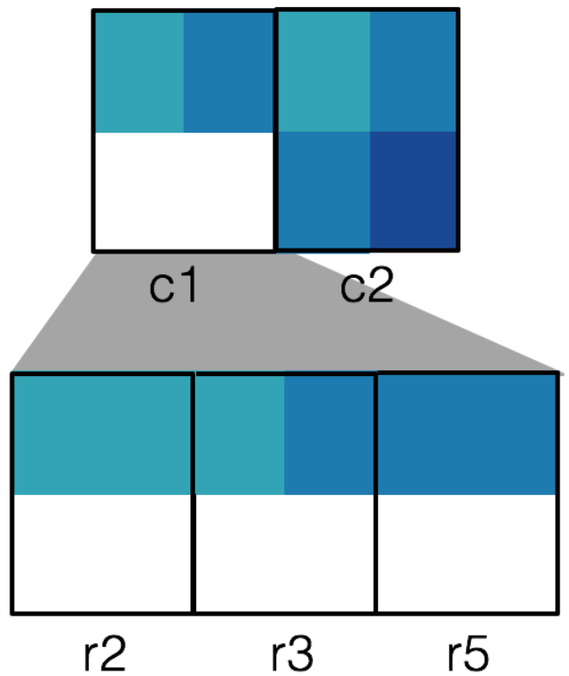
Apply *k*-means clustering with users best guess for *k*

Clustering | for navigation + exploration

A genomics example

Cluster averages

Individual matrices
Pan across all matrices
in the selected a cluster



Built an interactive visual interface to the clusters to enable exploration and understanding of a large dataset

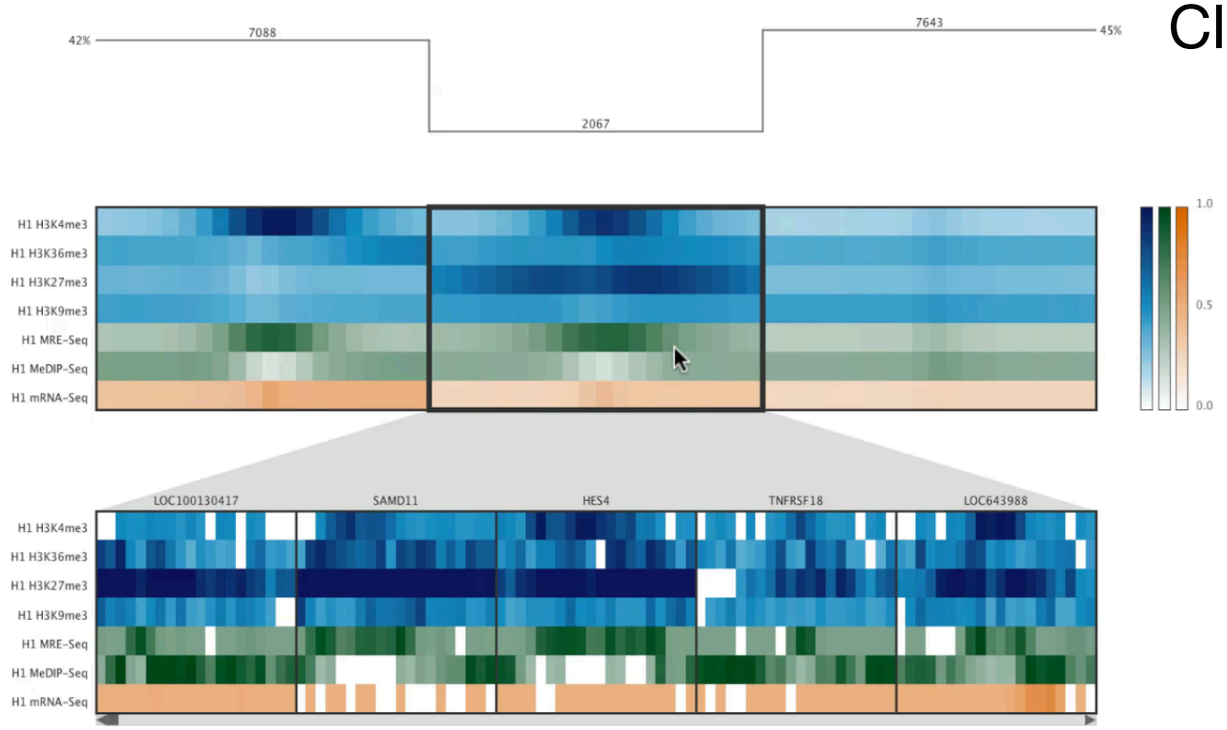
Allow users to interactively split or merge clusters

Clustering | for navigation + exploration

Cluster overview

Individual cluster members

Cluster sizes



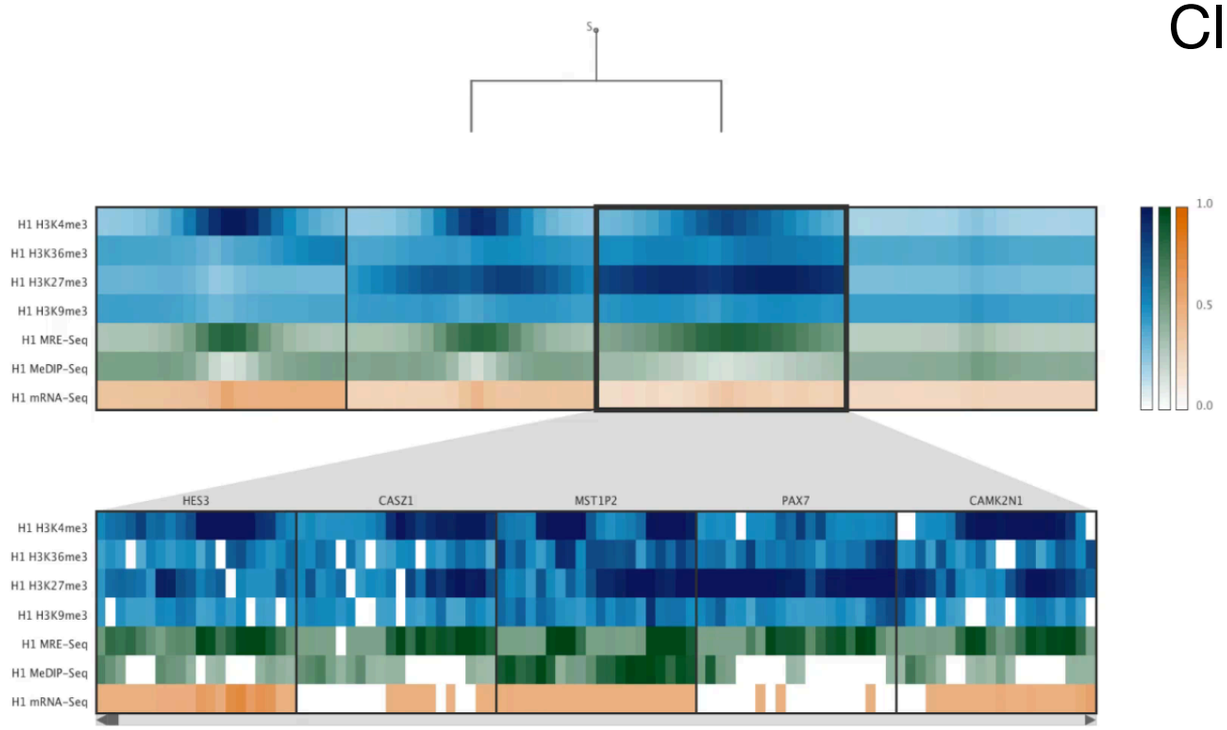
http://sparkinsight.org/

Clustering | for navigation + exploration

Interactively split 2nd cluster

Cluster split history

Cluster overview



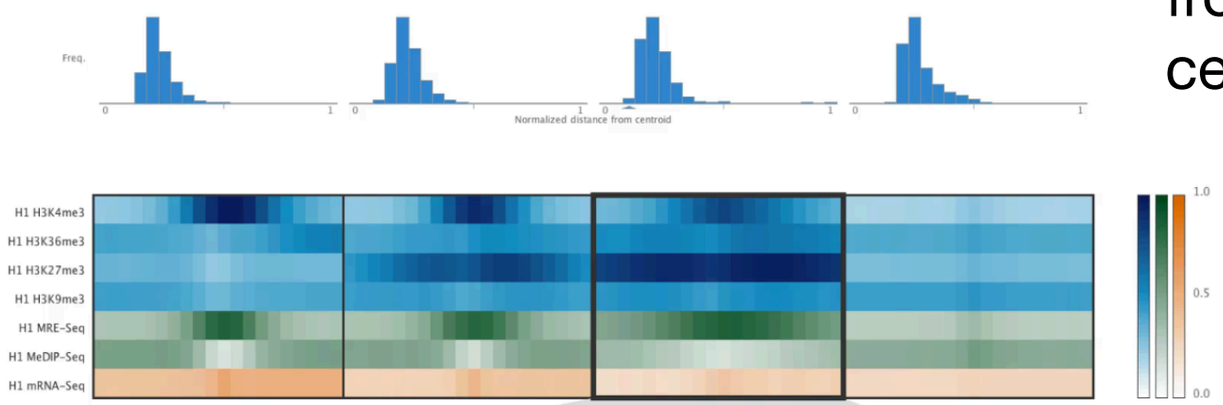
Individual cluster members

http://sparkinsight.org/

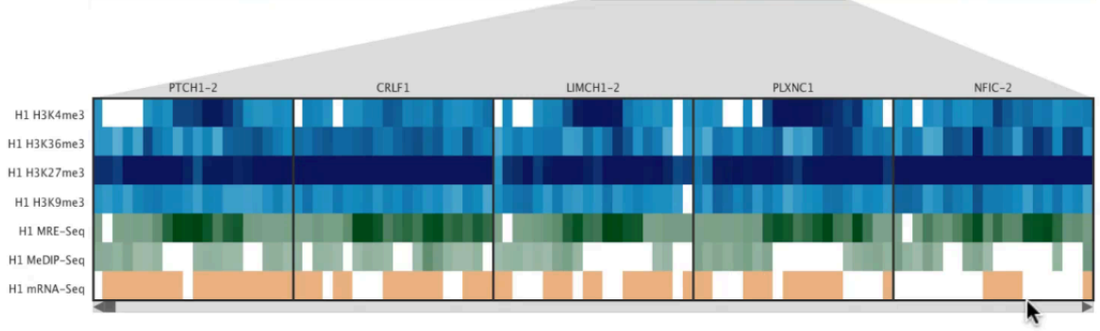
Clustering | for navigation + exploration

Distribution of member distances from the cluster centroid

Cluster overview



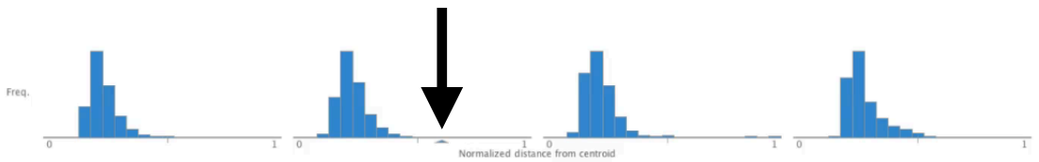
Individual cluster members



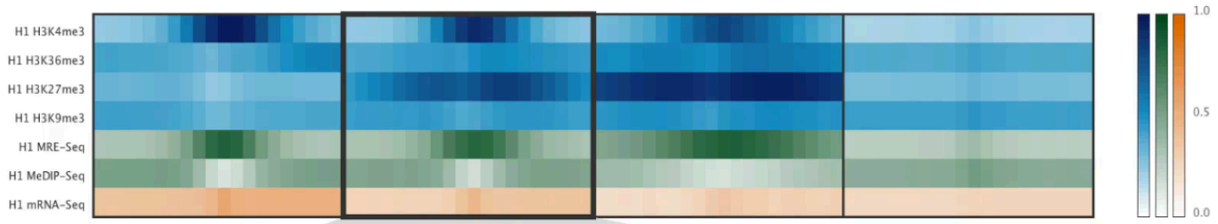
http://sparkinsight.org/

Clustering | for navigation + exploration

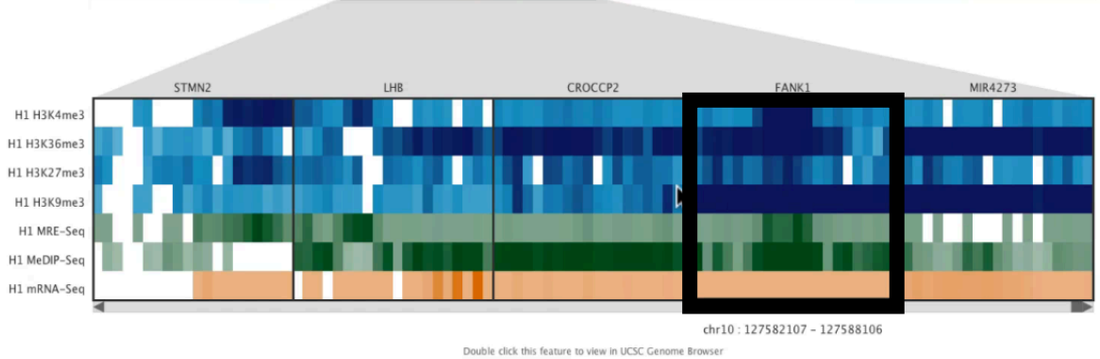
Distance of selected member to centroid is highlighted



Cluster overview



Individual cluster members



selected member

http://sparkinsight.org/

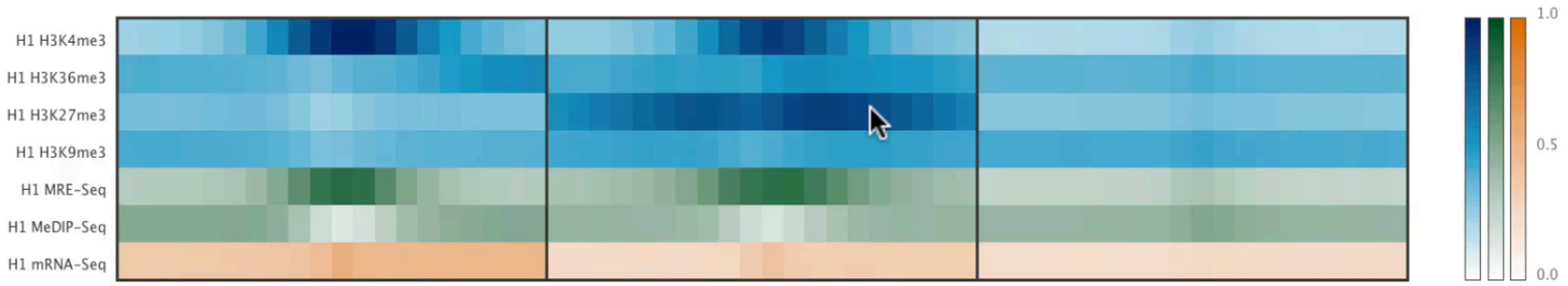
TSS

File Edit Help

GO COORD

Cluster Sizes (16798 regions clustered)

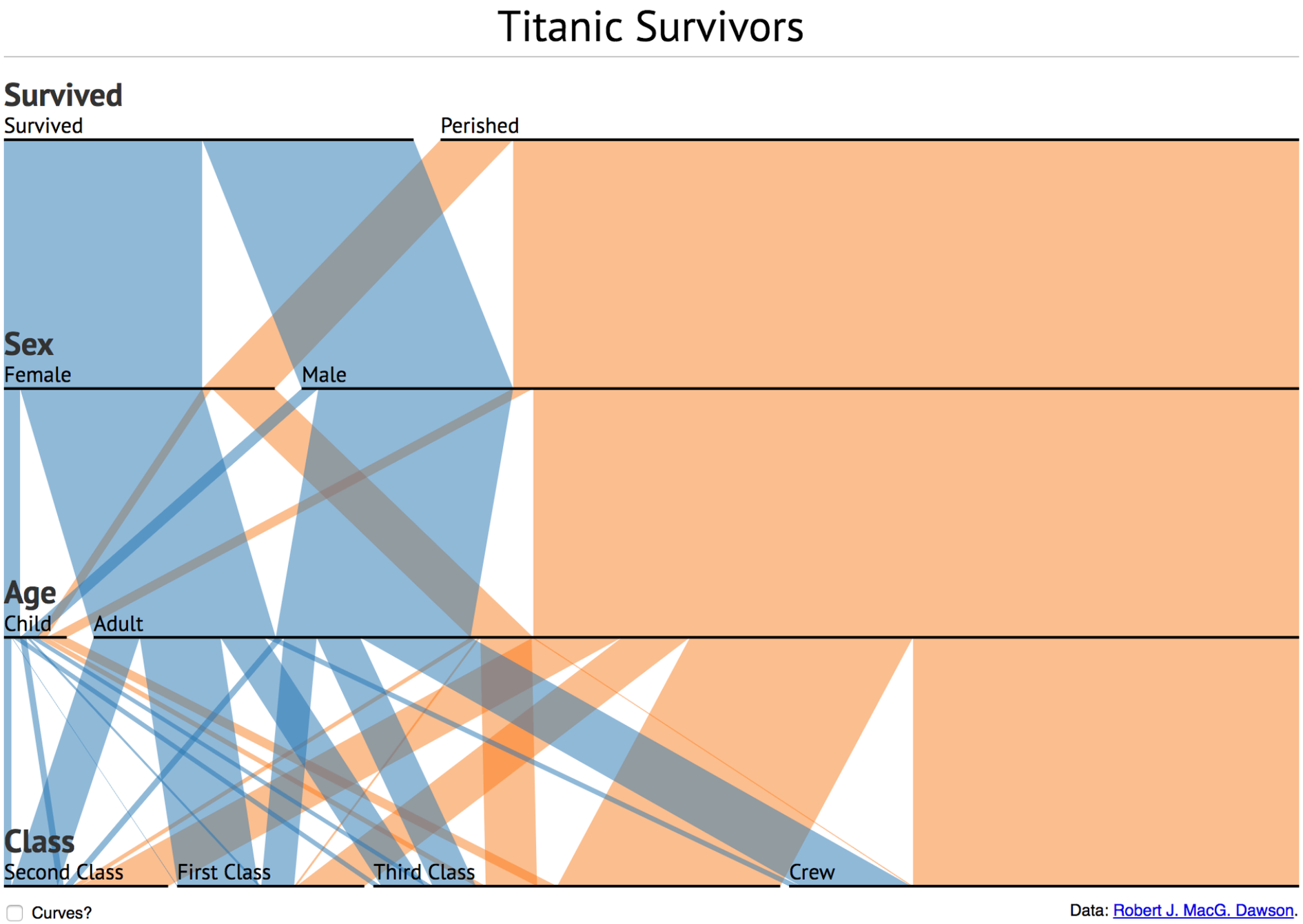
Cluster Profiles



http://sparkinsight.org/

Comparing different clusterings

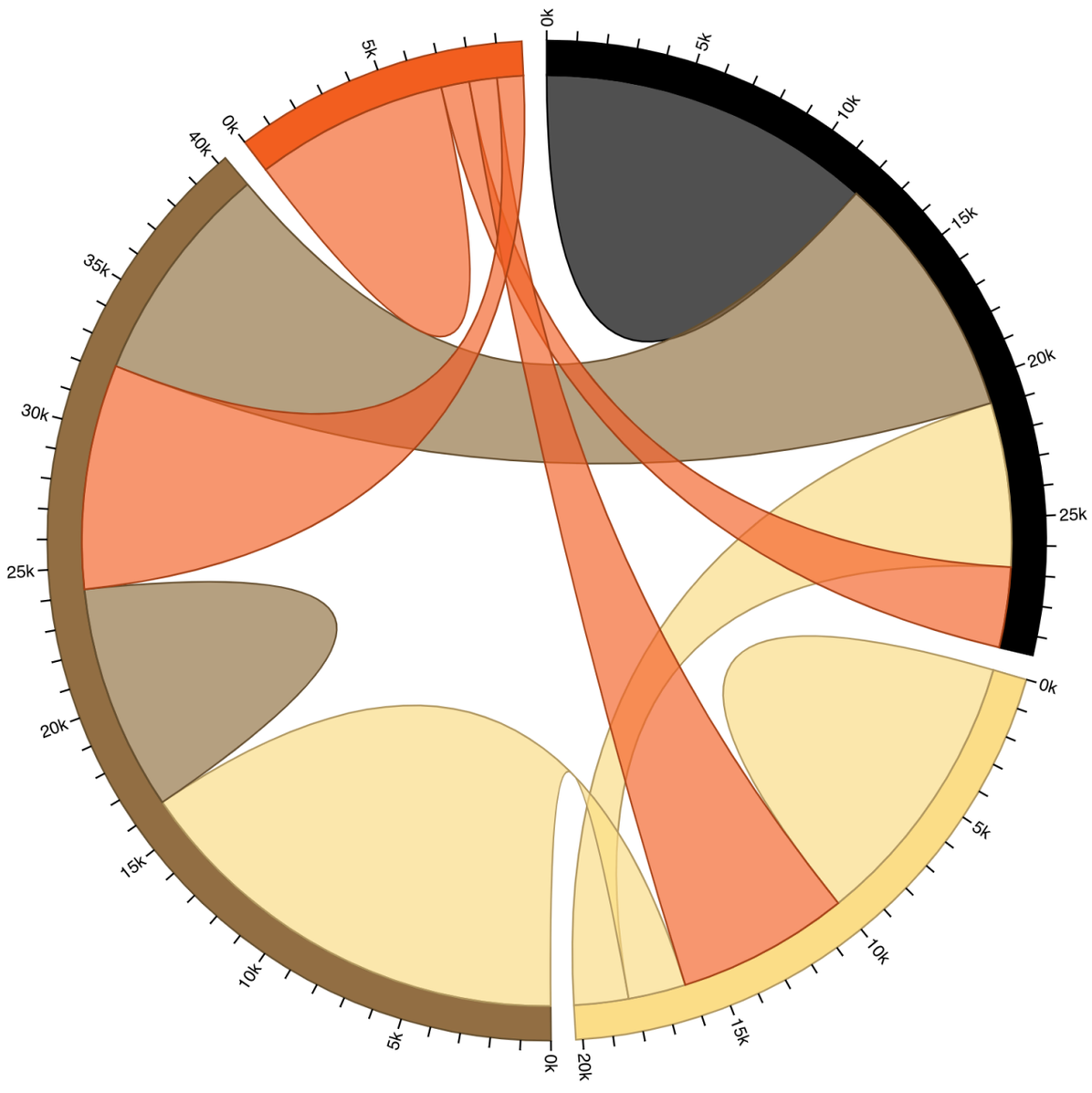
Parallel Sets



https://www.jasondavies.com/parallel-sets/

Comparing different clusterings

Chord diagram



https://beta.observablehq.com/@mbostock/d3-chord-diagram

Key points | Clustering

- **Hierarchical clustering** - Heatmaps + Dendrograms
  - Can be a powerful overview of a clustering
  - Order of rows and columns matters (often cluster both)
  - Interaction can help navigate (pan + zoom) and slice the dendrogram at different depths
- **Sequence clustering**
  - Can use model-based methods to cluster variable length sequences
  - Visualizing individuals in a cluster is more intuitive than visualizing the model's themselves
- **Making clustering interactive for exploration**
  - An initial grouping can serve as a useful navigational tool to explore data
  - Giving the user the power to change the grouping helps them explore it further and drill into clusters of interest
  - Sorting by distance from the centroid allows identification of outliers

Scalability

Research on big data visualization must address two major challenges:

perceptual and **interactive** scalability

Zhicheng Liu, Biye Jiang, Jeffrey Heer
inMens, EuroVis 2013

Perceptual scalability

What to do when you have more data points than pixels

You can go find more pixels

https://www.evl.uic.edu/documents/aurisano_bactogenie2015.pdf

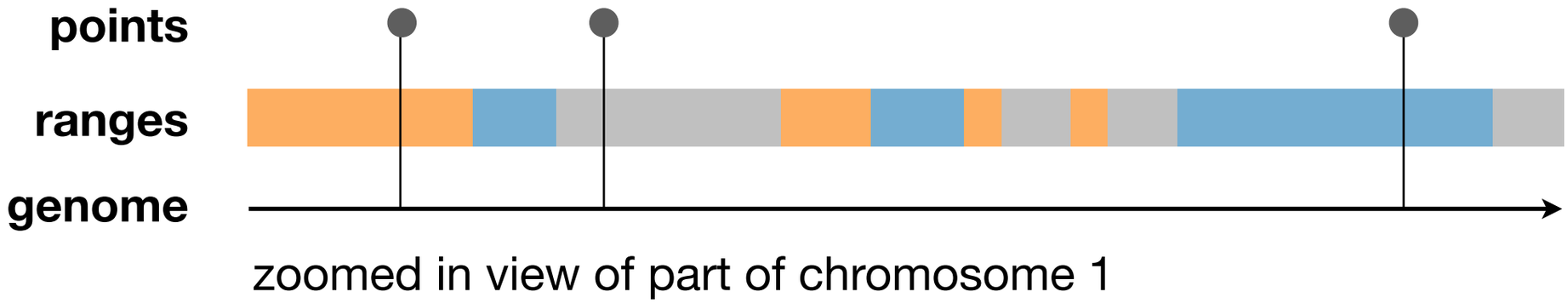
Large displays are not common

They come with their own design challenges

A real world example

Human genome has > 3 billion nucleotides

Scientists measure properties of **points** and **ranges** along that sequence



zoomed in view of part of chromosome 1

Store as tabular data

Points

chr	start	end	probability
1	13,205	13,206	0.91
1	20,998	20,999	0.88
1	68,648	68,649	0.76

Hundreds of thousands up to millions of points

Ranges

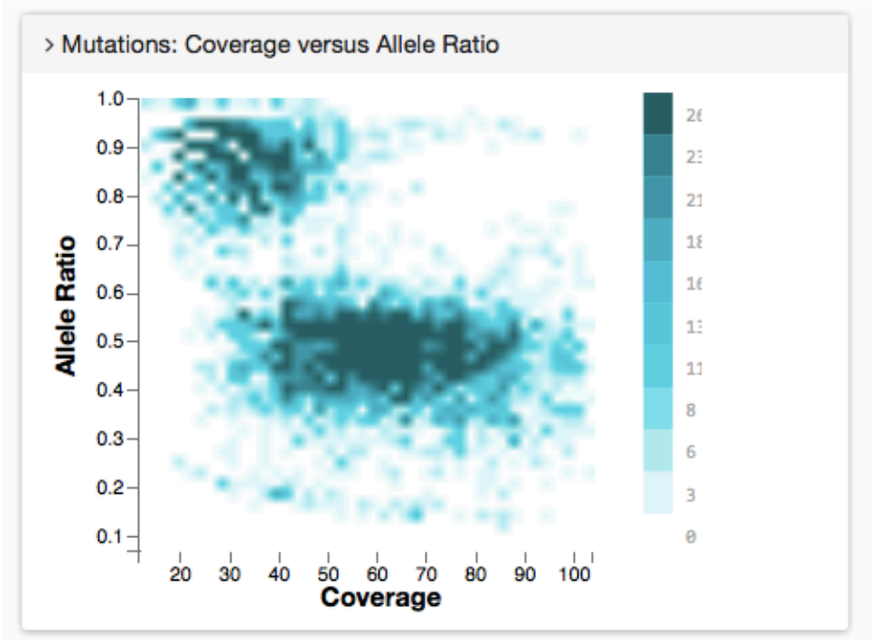
chr	start	end	category	probability
1	0	15,824		0.91
1	15,825	21,190		0.88
1	21,191	40,983		0.95

• • • rest not shown

Up to tens of thousands of ranges

Scatter plot of 100,000+ points

Binned scatterplot



Approach

- Apply a rectangular grid to a scatterplot
- Count number of points in each rectangle
- Display counts on colour scale (heatmap) instead of the raw points

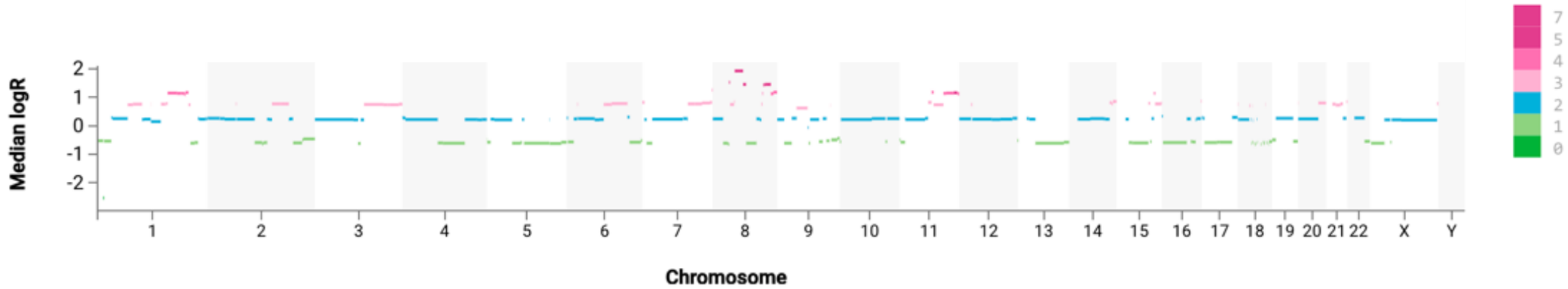
Benefits

- Solves over-plotting problem when have too many points
- Can be faster to render

Montage System

BC Cancer

Can’t plot them all | hint at what’s missing



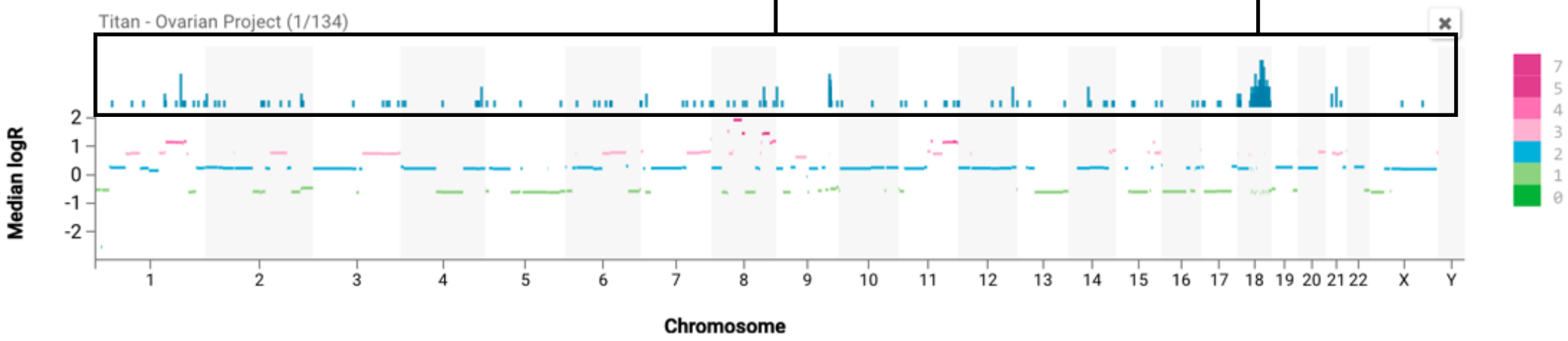
>3 billion nucleotides

Want to display in <1000 pixels

Can’t plot them all | hint at what’s missing

As you zoom into this peak, small features become visible (peak tells you where to look)

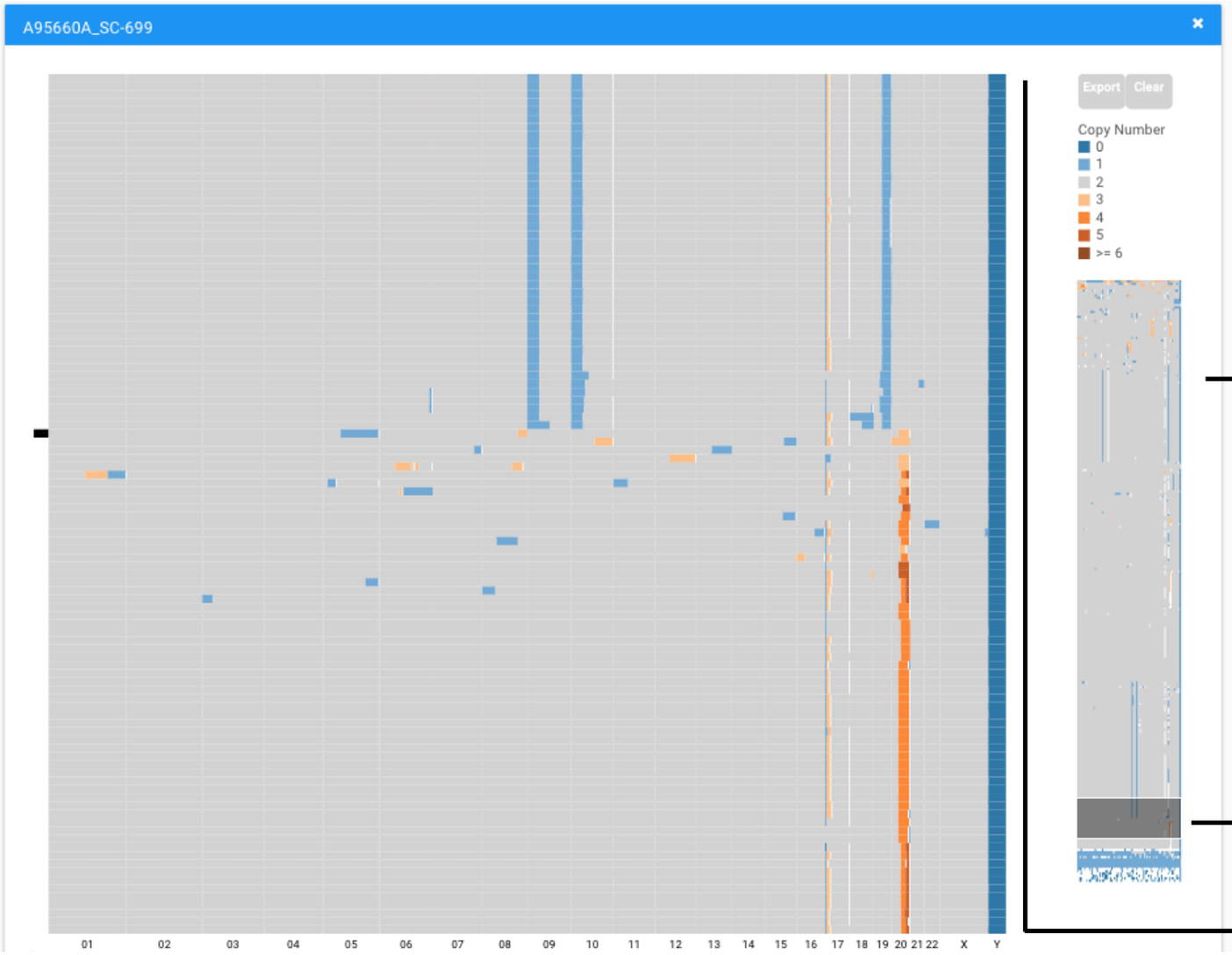
Histogram of how many **features are too small** to be displayed in the current view



>3 billion nucleotides

Want to display in <1000 pixels

That’s just one genome, what about many



Sequence data for thousands of genomes (actually genomes from single cells!)

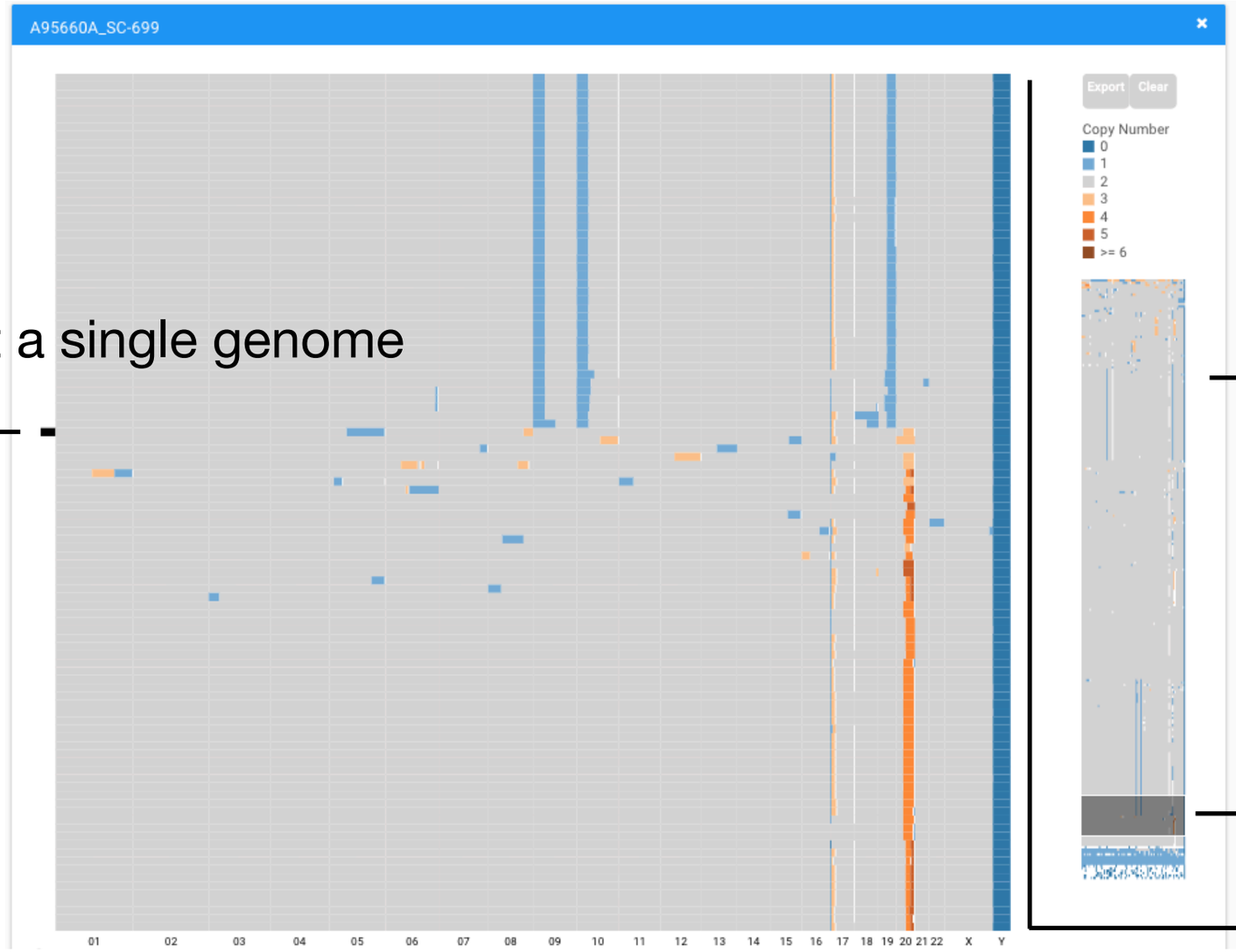
Minimap

row = single genome

selection in minimap

zoomed in view of selection

That’s just one genome, what about many



select a single genome

Sequence data for thousands of genomes (actually genomes from single cells!)

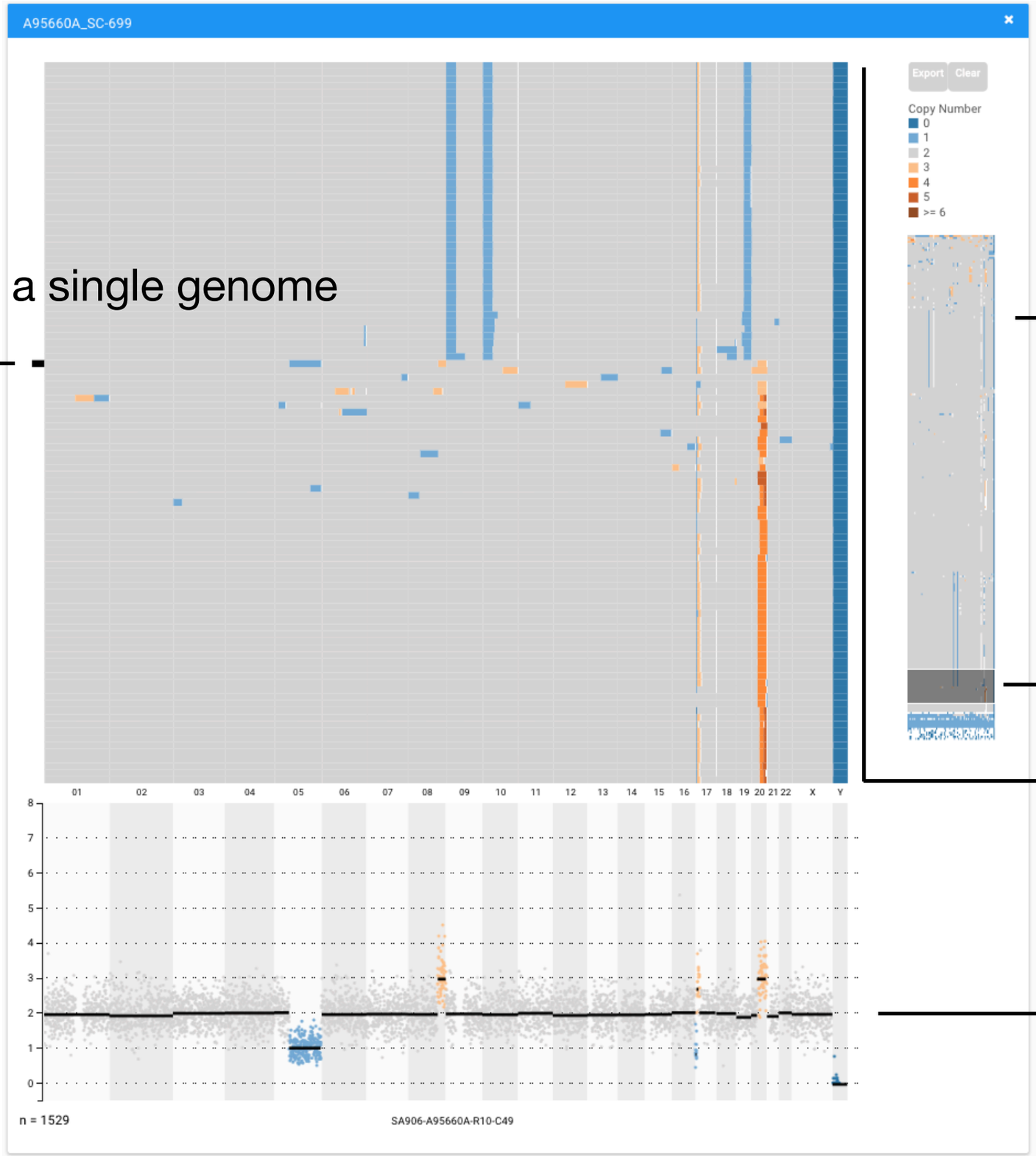
Minimap

row = single genome

selection in minimap

zoomed in view of selection

That’s just one genome, what about many



Sequence data for thousands of genomes (actually genomes from single cells!)

select a single genome

Minimap

row = single genome

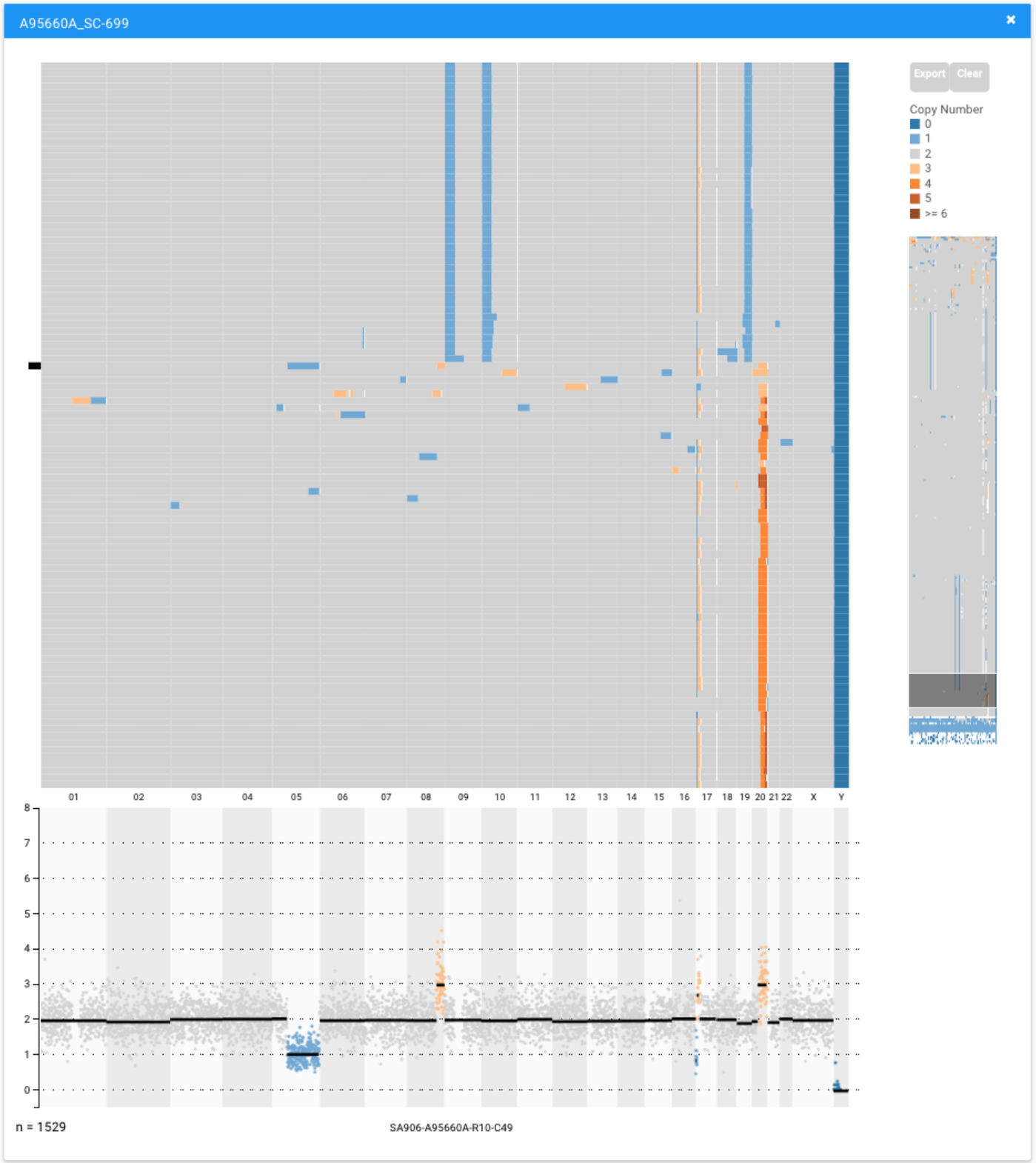
selection in minimap

zoomed in view of selection

black lines = ranges

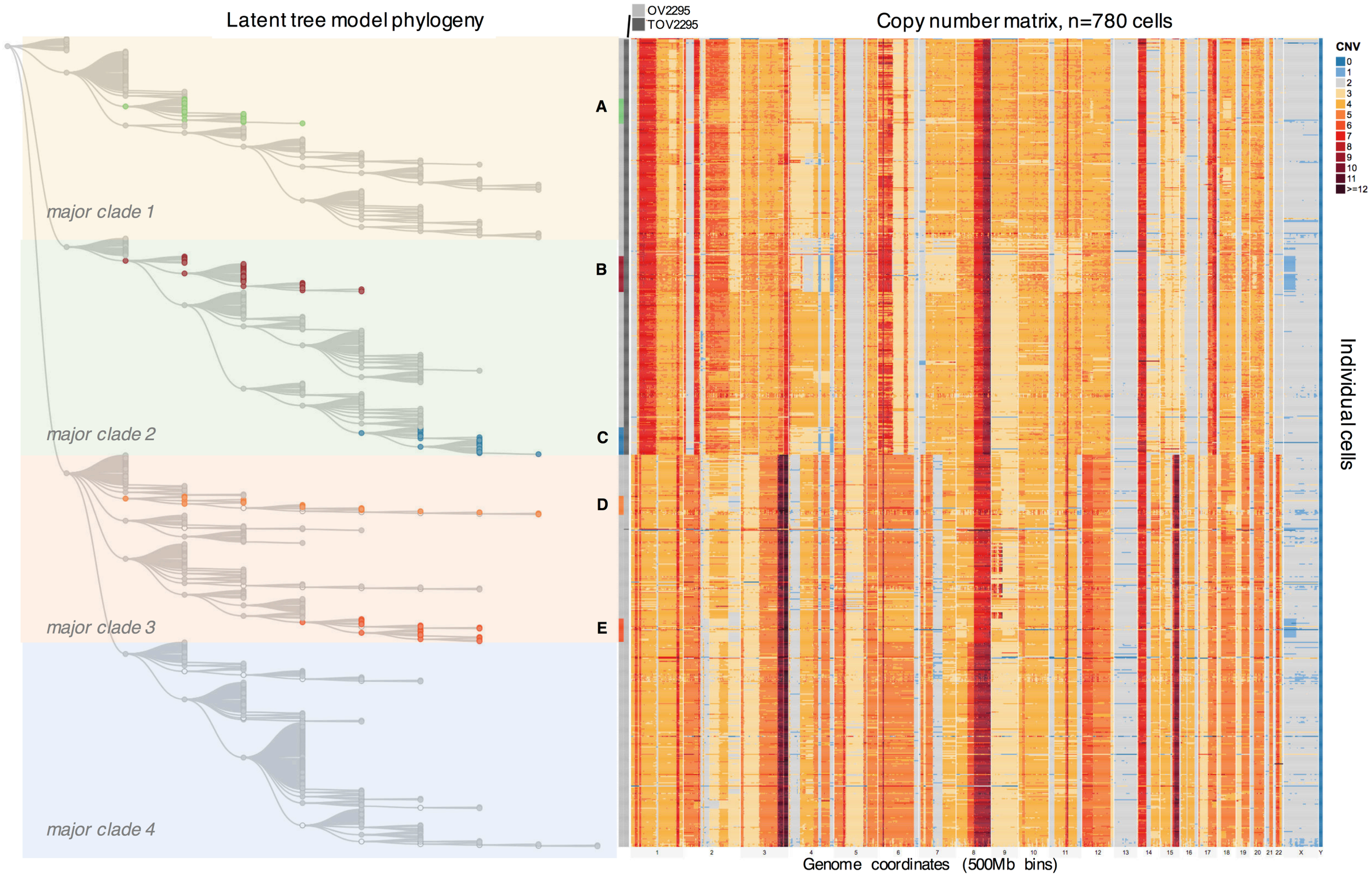
additional points are a raw form of data used to predict the ranges

That’s just one genome, what about many



Minimap displays a sample of genomes (cannot show all)

Could have considered sampling the supporting points in the detailed view (dots)



Later introduced a phylogenetic tree to show relationships between the genomes

As grew to millions of genomes, explored methods of summarizing the tree

Sohrab Shah lab
BC Cancer

Interactive scalability

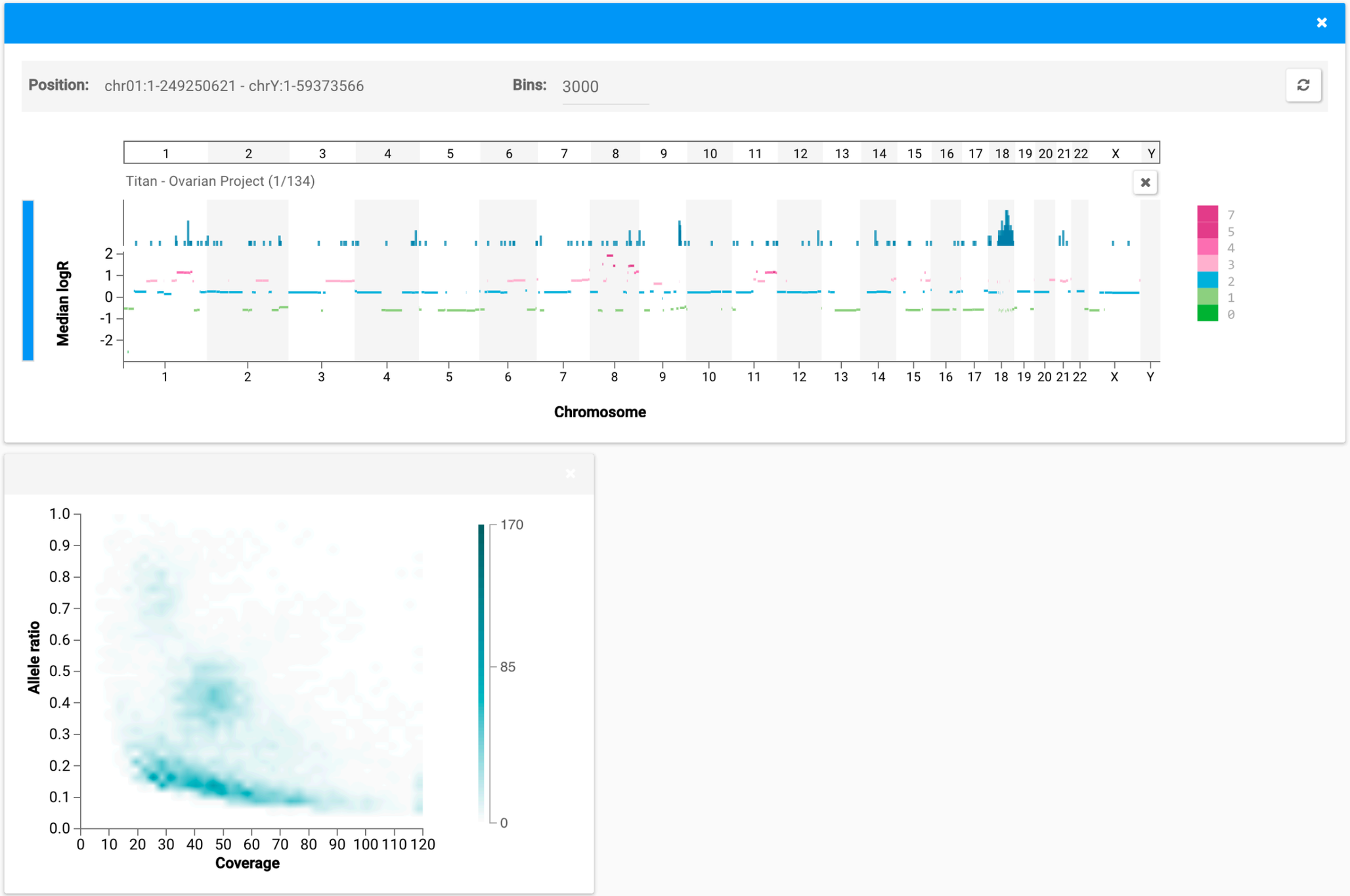
How to make your visualization of a large dataset responsive

Responsiveness

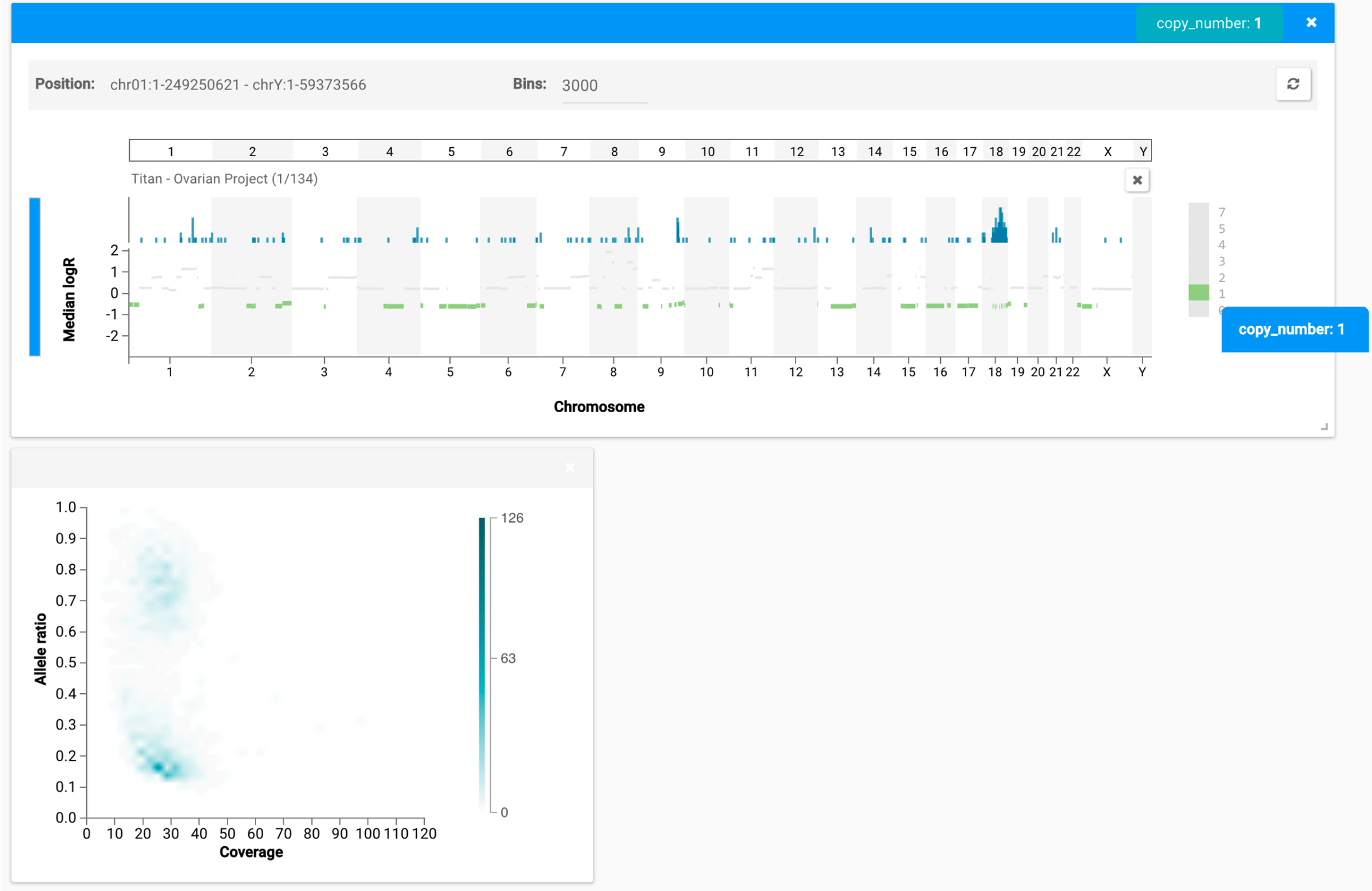
Visual feedback: three rough categories

- **0.1 seconds: perceptual processing**
  - subsecond response for mouseover highlighting
- **1 second: immediate response**
  - response after mouseclick, button press
- **10 seconds: brief tasks**
  - Multi-second operations should show progress or processing indicator, ideally with option to cancel

Linked views



Linked views

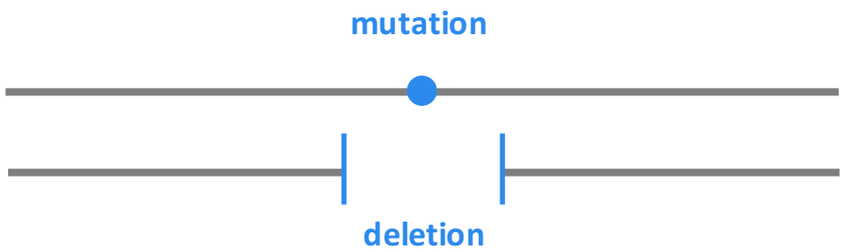


Linked views

a chromosome (2 copies)

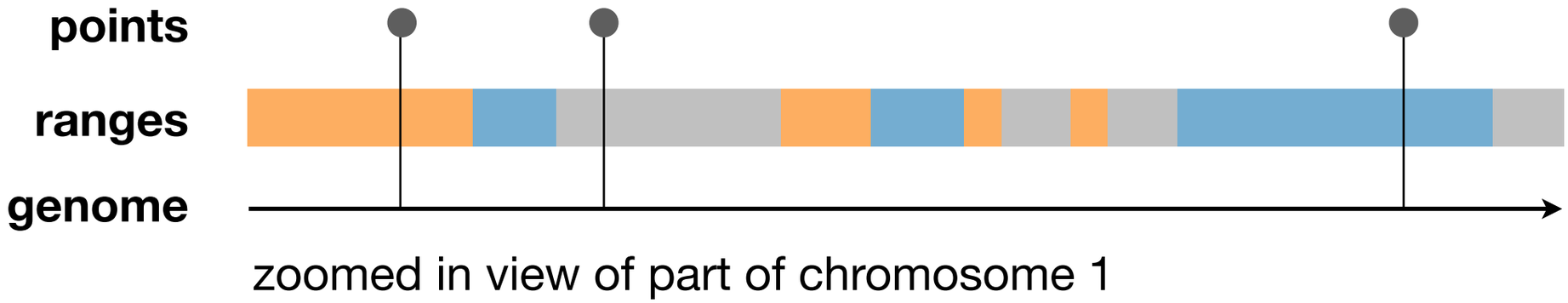
Linked views

Linked views

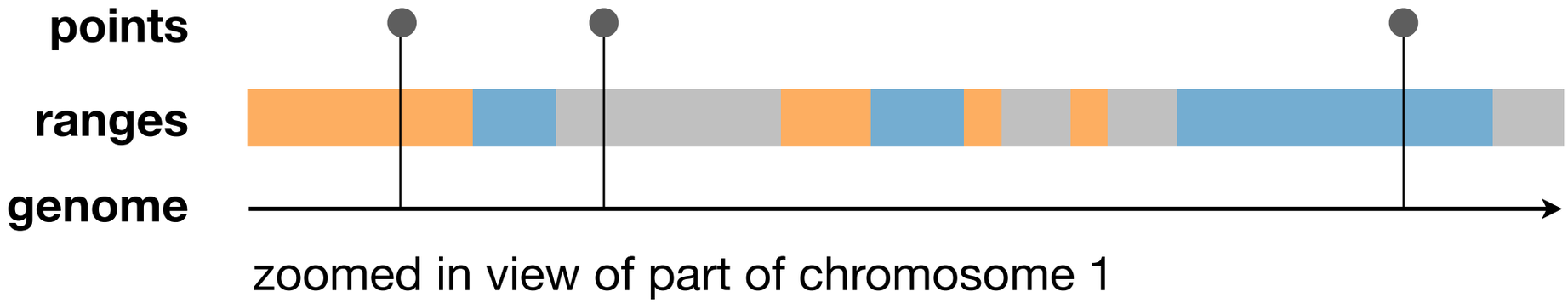


a chromosome (2 copies)

Linked views



Linked views



Points

chr	start	end	probability
1	13,205	13,206	0.91
1	20,998	20,999	0.88
1	68,648	68,649	0.76

Ranges

chr	start	end	category	probability
1	0	15,824		0.91
1	15,825	21,190		0.88
1	21,191	40,983		0.95
• • •			rest not shown	

Linked views

- **Points** and **ranges** are in separate tables
- To find the overlaps between the selected ranges and points on-the-fly would be too slow for potentially millions of items
- Must pre-compute the overlaps

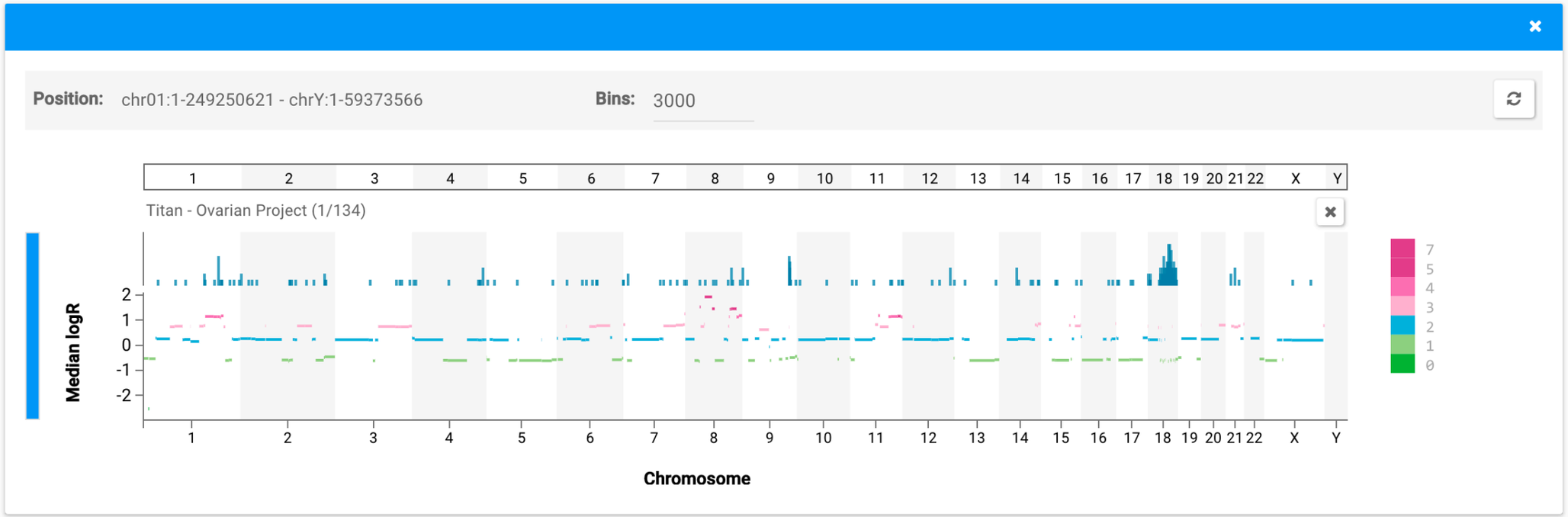
Points

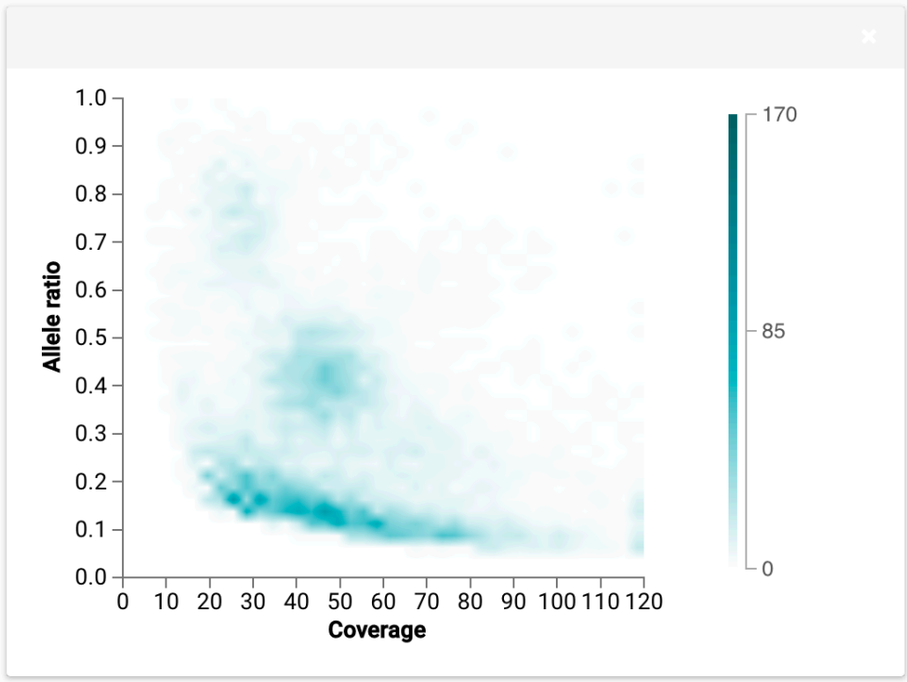
chr	start	end	probability
1	13,205	13,206	0.91
1	20,998	20,999	0.88
1	68,648	68,649	0.76

Ranges

chr	start	end	category	probability
1	0	15,824		0.91
1	15,825	21,190		0.88
1	21,191	40,983		0.95
• • • rest not shown				

Linked views

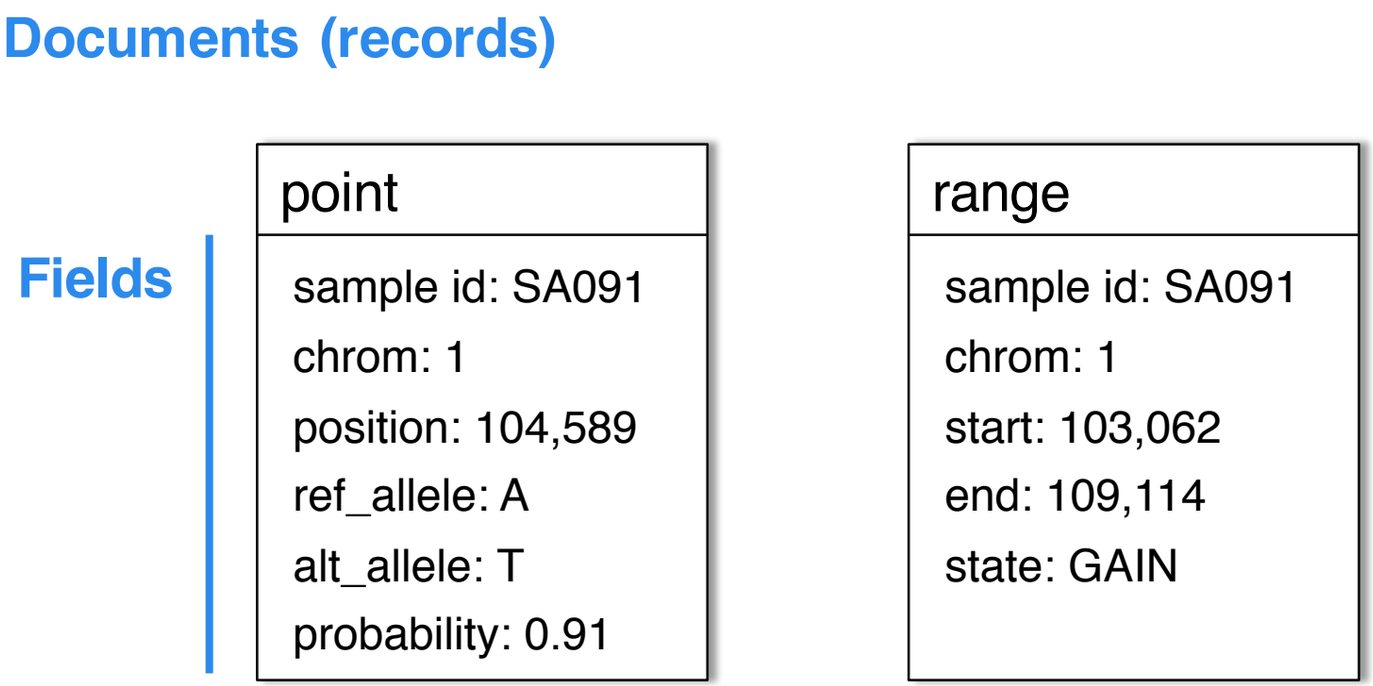




To support bi-directional linking, need to store:

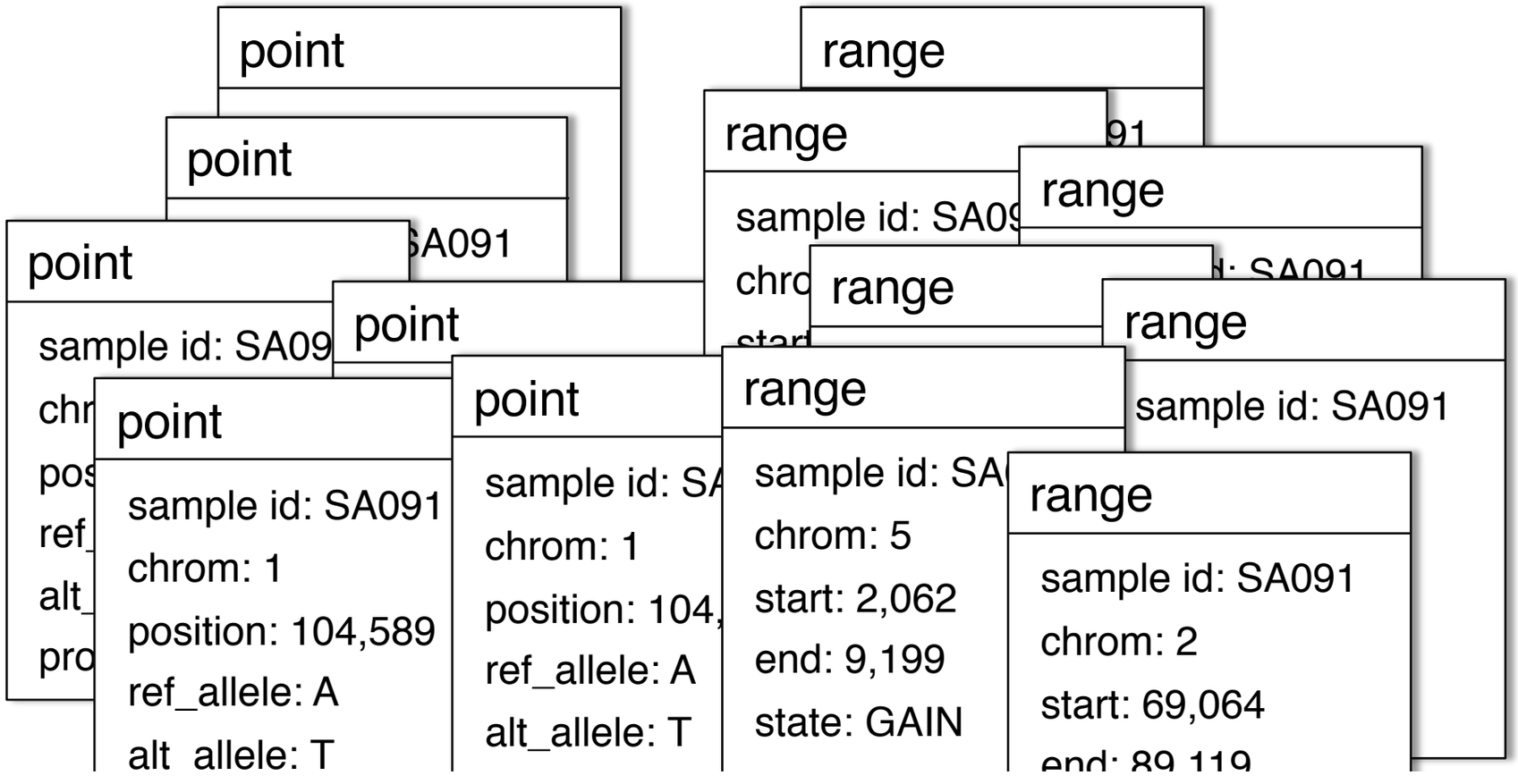
- overlapping points for each range
- overlapping ranges for each point

Used Elasticsearch



Used Elasticsearch

Index



Denormalized the data

point

sample id: DAH177
library id: A32068
type: mutation
chrom_number: 3
start: 139290475
end: 139290475
probability: 0.91
…
events: (
(type: range,
chrom_number: 3
start: 111191194
end: 154780786
copy_number: 3
…
)
)

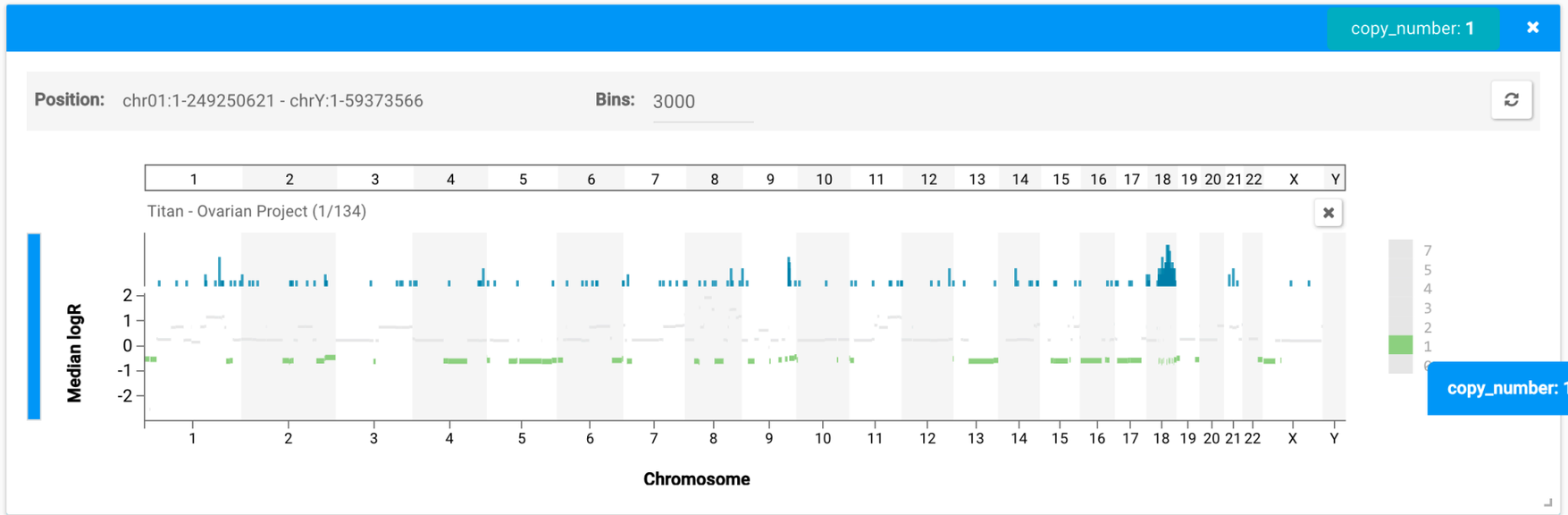
range

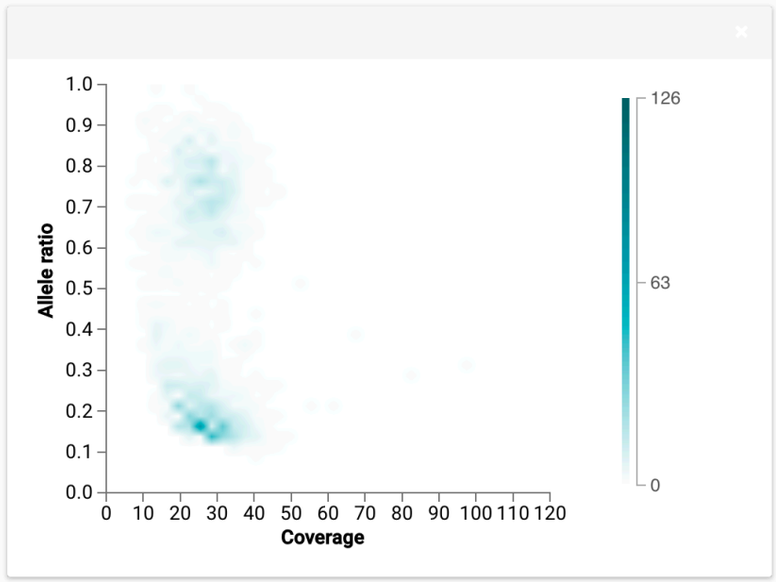
sample id: DAH177
library id: A32068
type: segment
chrom_number: 3
start: 111191194
end: 154780786
copy_number: 3
…
events: (
(type: point,
chrom_number: 3
start: 139290475
end: 139290476
probability: 0.91
…
)
)

- Copy the information about the overlapping points/ ranges into each overlapping item's document
- Exploit search engine's very fast retrieval of terms

Linked views across different data types

selection: copy_number = 1





select all mutations
where events.copy_number = 1

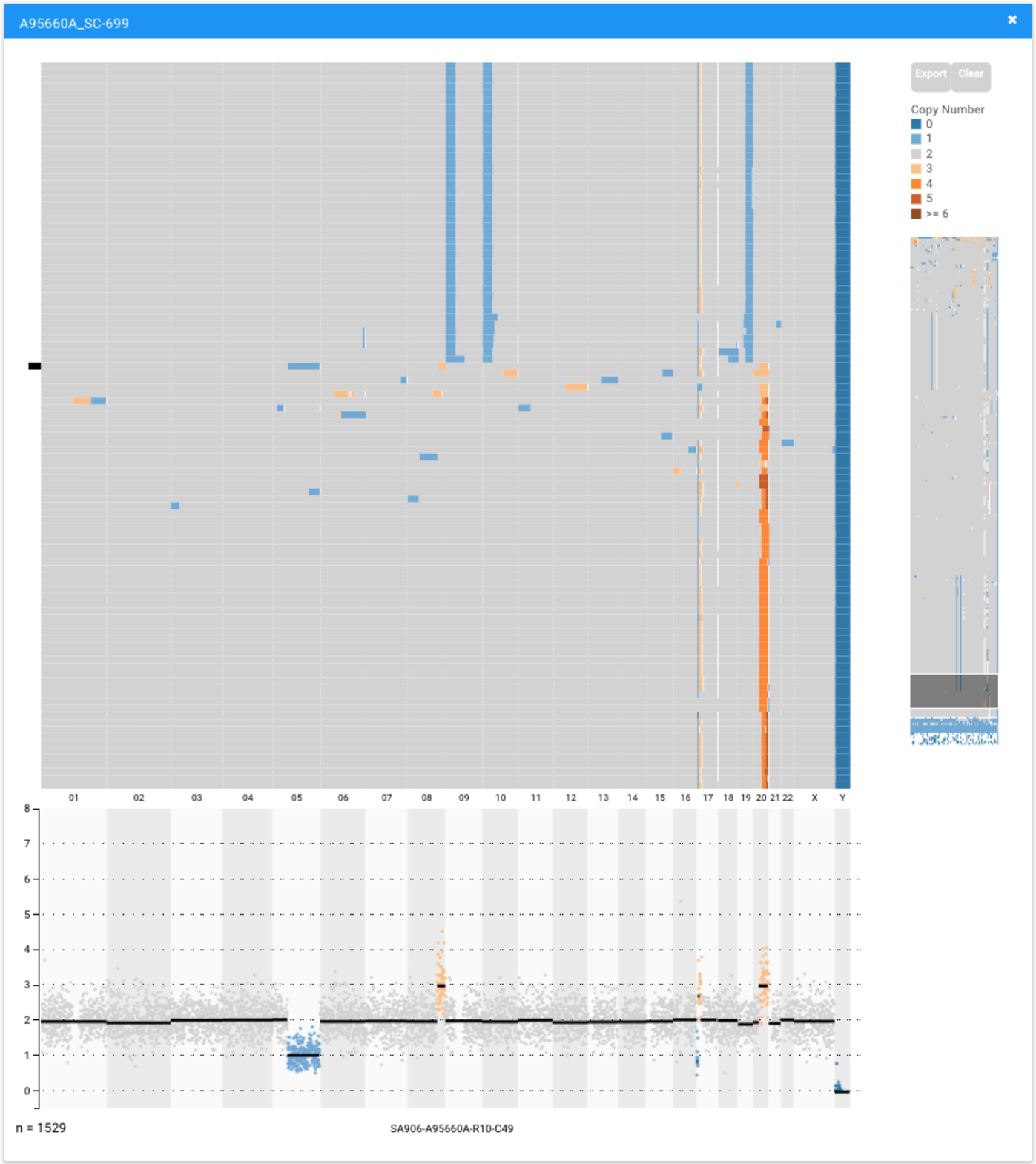
Montage System

BC Cancer

Elasticsearch built-in aggregations ensured binning of millions of points remained responsive

Pre-compute / transform data so that query speed is optimized (design your data store to serve the queries in your interface)

Other tips and tricks



Introduced filters to remove sub-pixel ranges from being returned in our query results

- can't see them anyway
- smaller query result faster transfer over the network

Draw detailed view using Canvas not SVG (better performance on millions of items)

Cached supporting point data so detailed view immediately responds when a single genome is selected above

Key points | Scalability

- **Perceptual scalability**
  - What to do when you have more data points than pixels
  - Be aware of sub-pixel elements
    - Find another way to represent them, or hint that they have been removed
  - Binned scatterplots are a great way to deal with over-plotting
  - Sampling in an overview (mini-map) and details (on-demand)
- **Interactive scalability**
  - How to ensure a responsive interface despite millions of data points
  - Pre-compute / transform data so that query speed is optimized (design your data store to serve the queries in your interface)
  - Pre-cache values where possible so they are immediately available on demand
  - Replace SVG with Canvas when element count grows large

Office hour today 12-1pm

ESB 1045