Elasticsearch Visualization

co-op term presentation

Claire Barretto

Over the past 4 months

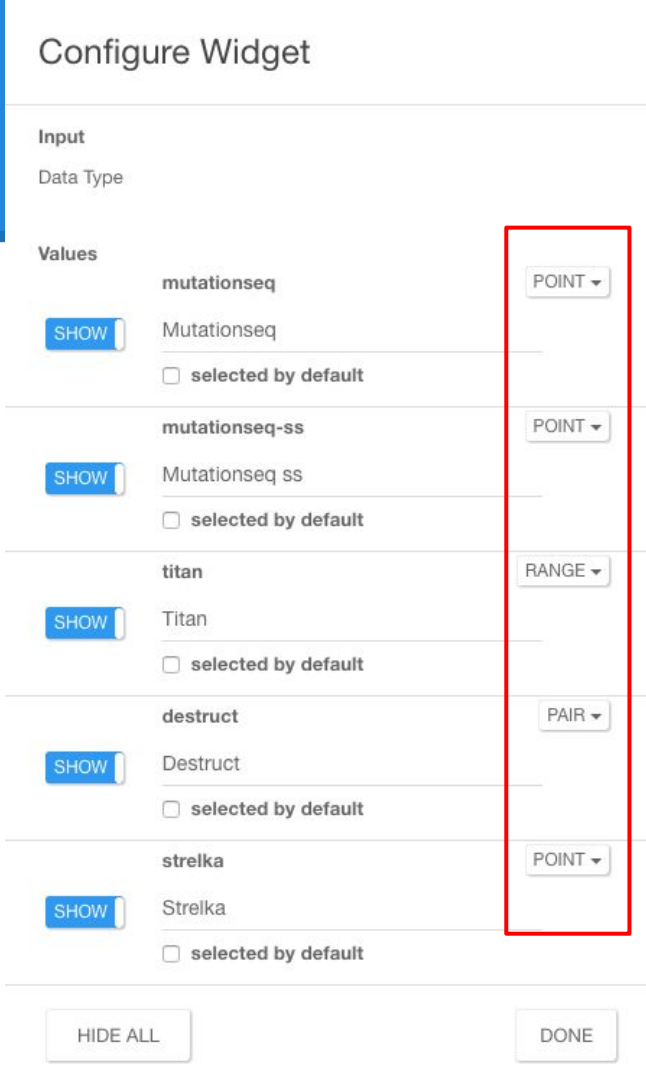
- General genome-wide plot
  - Made up of three types of views:
    - Generalized **range**-based view (eg. Titan)
    - New **pair**-based view (eg. Destruct)
    - New **point**-based view (eg. MutationSeq, MutationSeq SS)
- Stackable views (tracks)
- Zooming/Panning
- Selection

Genome Wide Plot

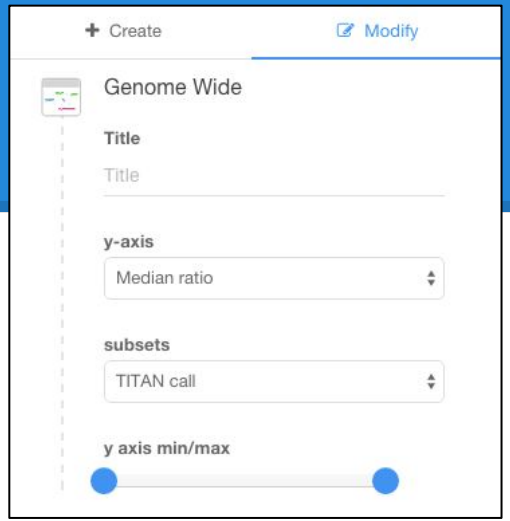
Most plots work with all data types, but for the genome-wide plot, we want to generate specific plots for certain types of data. This requires knowledge of the types of records in a data type

Configuration panel

- Data types are configured to indicate what its record types are
- Uses start and end position to determine if a record is a point, range or pair

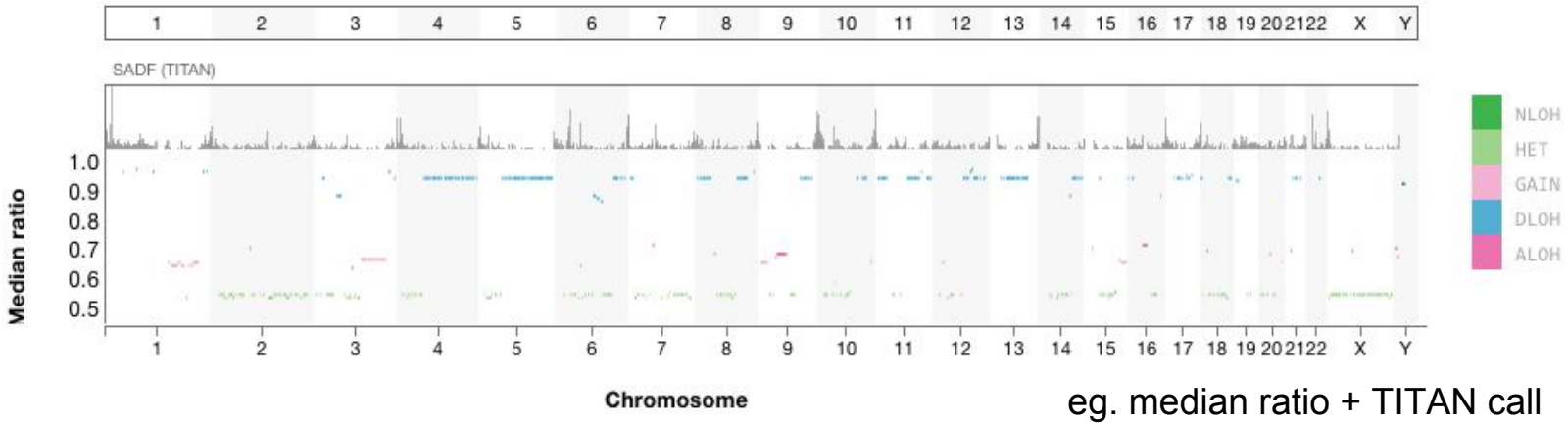


Generalized Range View



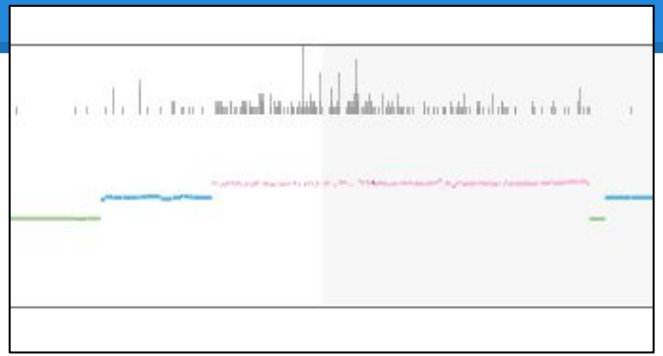
Existing copynum plot was hard coded to display Median logR + copy number. Wanted this plot to also be dynamic like other plots.

- Generalized so that it takes in any numerical field as its y-dimension, any categorical field as its subset



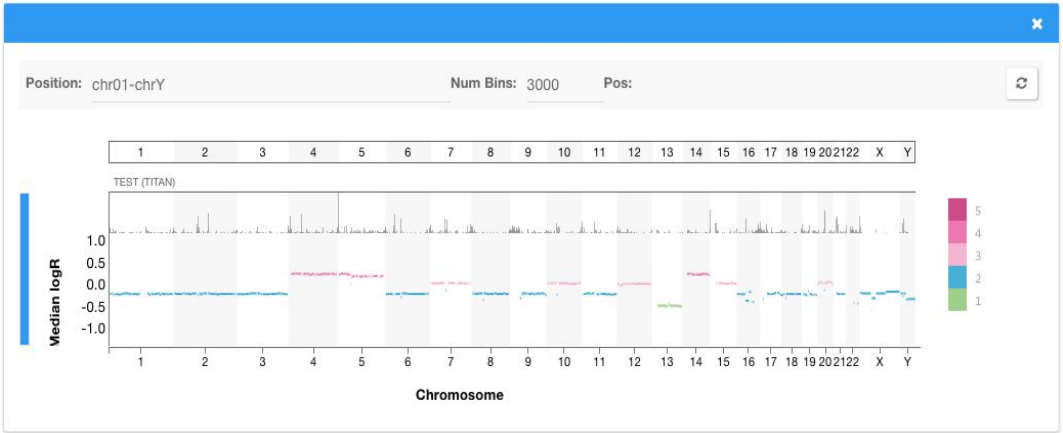
eg. median ratio + TITAN call

Optimizing Rendering

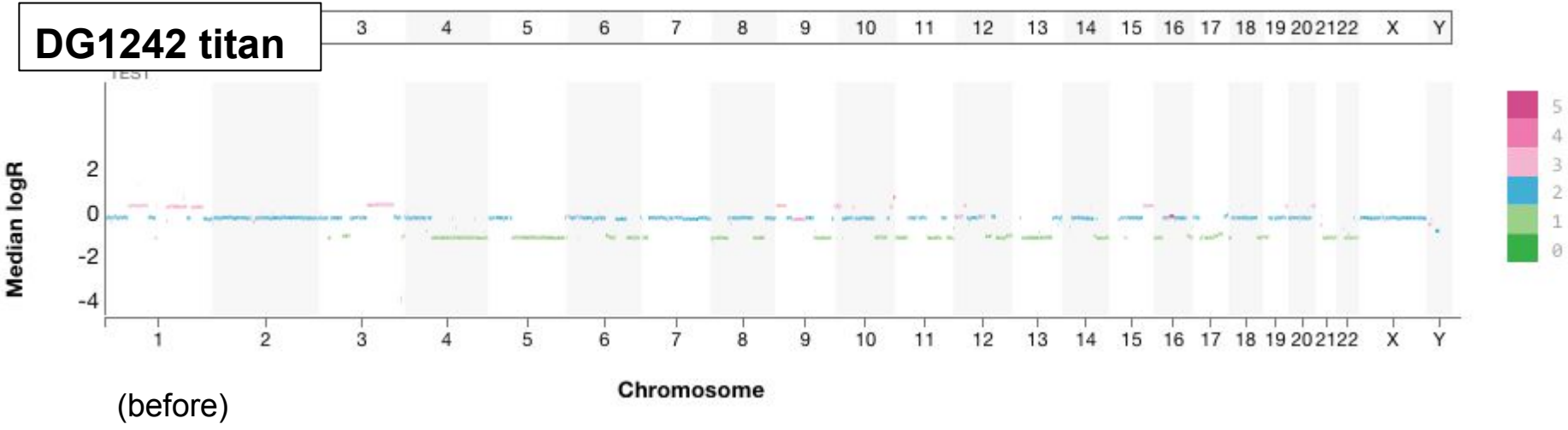


Wanted to reduce the number of elements plotted on the DOM, while also preserving potentially interesting smaller points

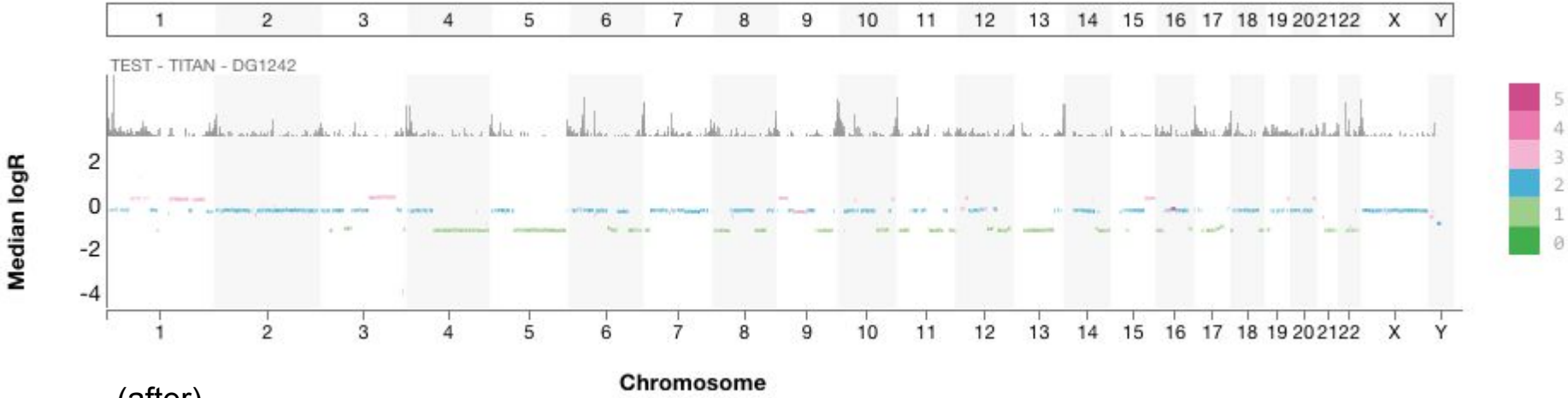
- Apply minimum threshold (length)
- Zoom in to view finer points
- Use density histogram to see areas that have many smaller points



DG1242 titan



(before)



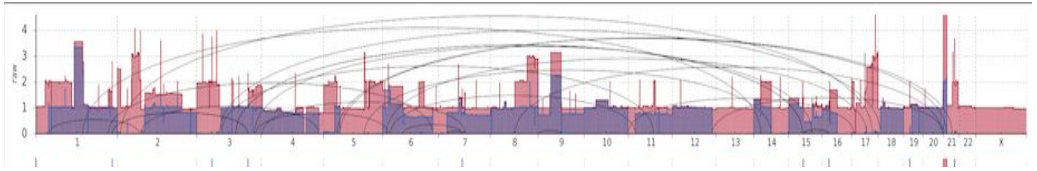
(after)

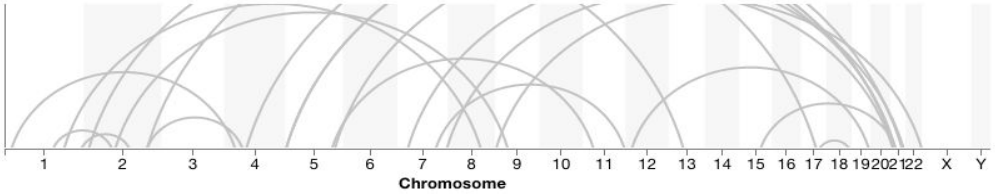
Pair View Initial Concept

Experimented with different techniques to visualize paired data. Started with arcs.

Limitations

- Points that are interesting are close together, hard to see
- Hard to follow trajectory
- Large arcs are visually dominant, but are not necessarily more important

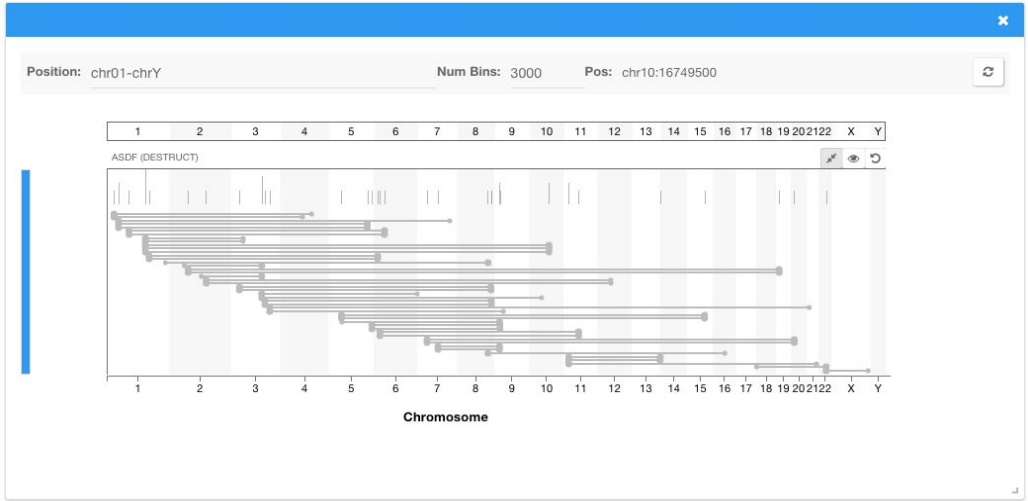




Pair View

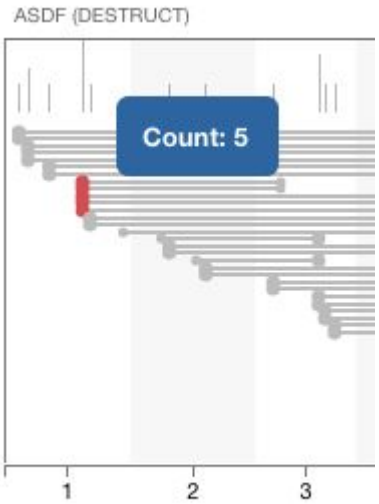
New rearrangement plot

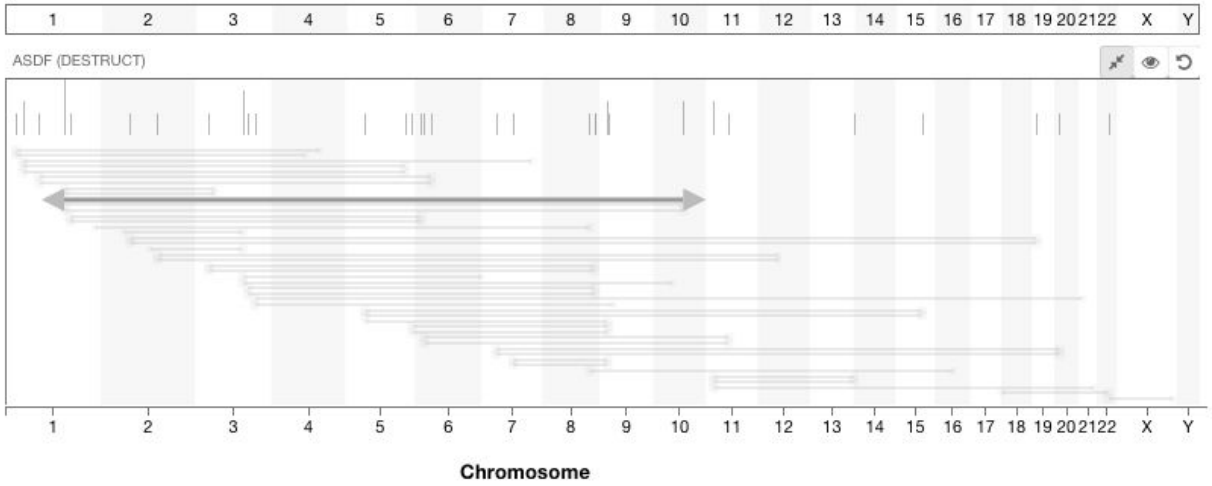
- Uses lines to visualize where rearrangement events occur
- Sorted by start position to see areas where many breakpoints fall
- Easier to follow path, non-overlapping
- Long lines are less dominating because lines are equally spaced vertically



Pair View Features

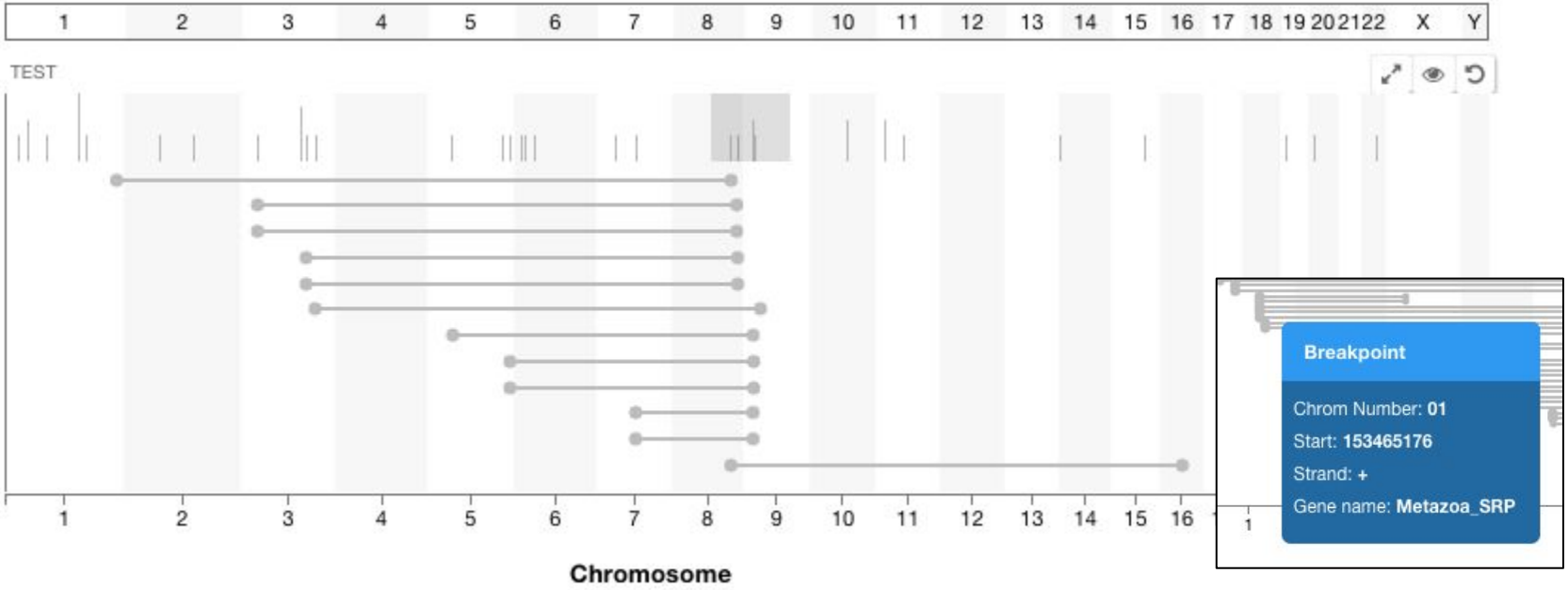
- Discover hotspots - breakpoint density histogram show areas where multiple breakpoints fall
- Identify reciprocal events using strand info, displayed on hover





Pair View Features

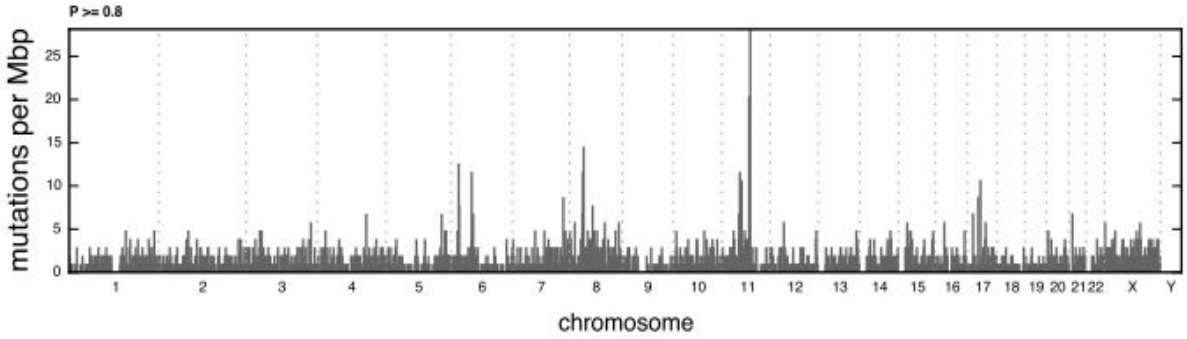
- Selection of bins - view all breakpoints that fall within the selected bins

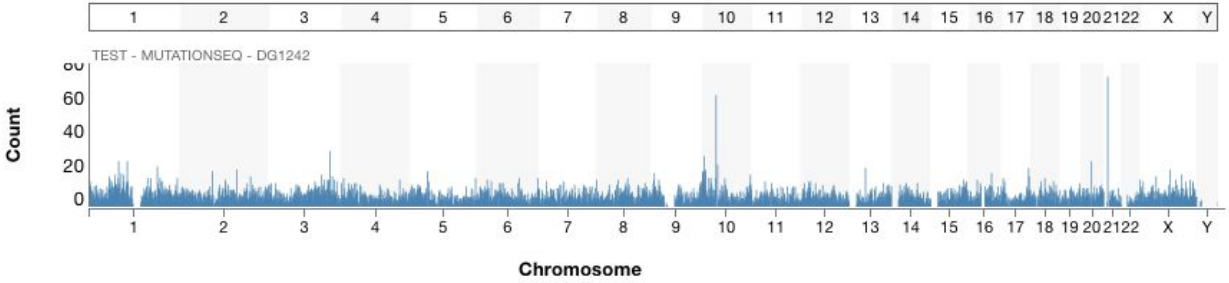


Point View

Incorporated mutation density chart (like the one in MutationSeq portraits) into interface

- Shows mutations per bin to identify hotspots/ peaks.
- Adjustable number of bins

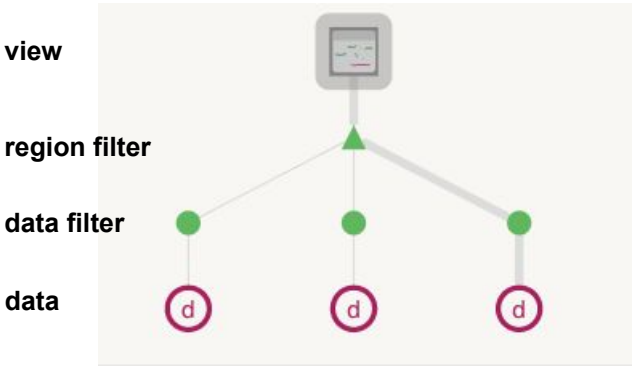




Tracks

Wanted a way to easily compare two or more plots next to each other

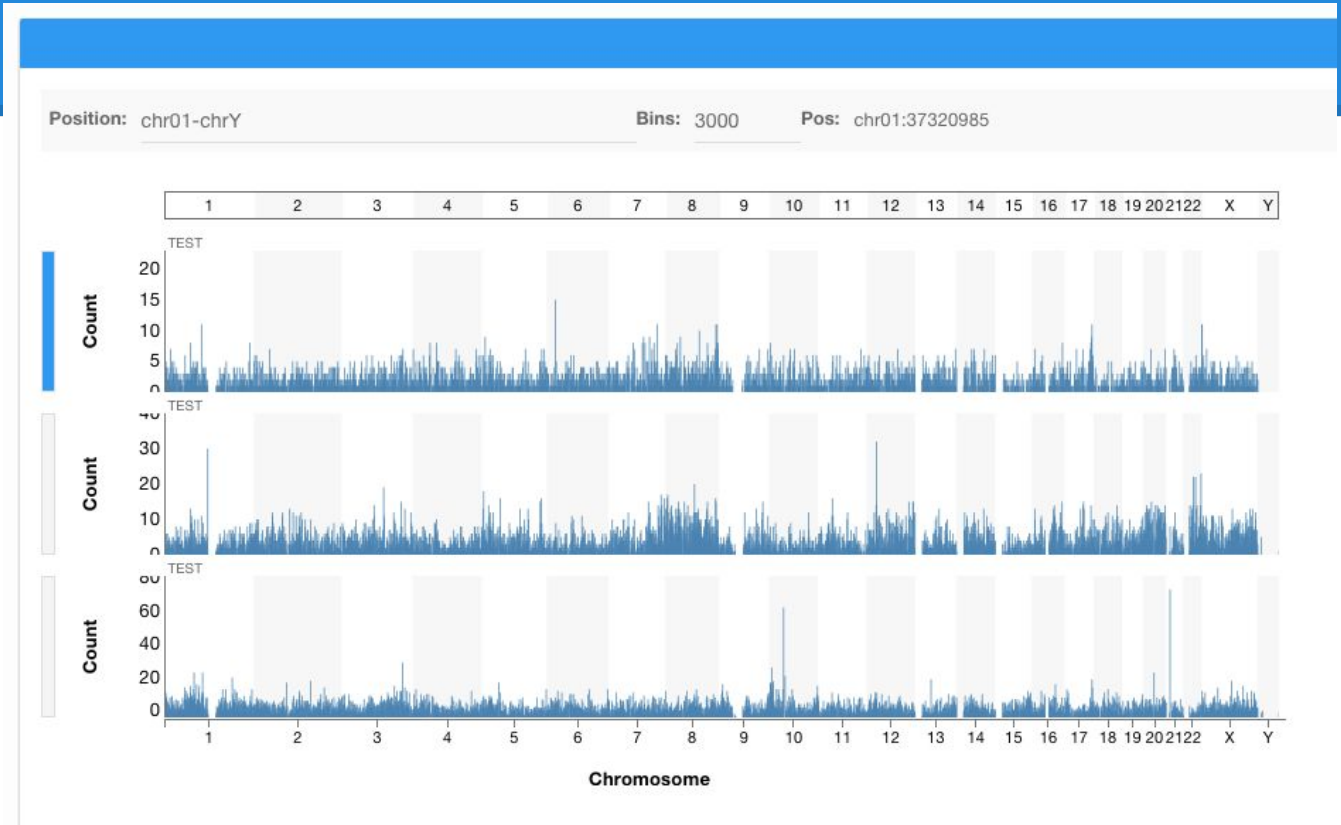
- Multiple data sources to one view
- Unique data filters for each track
- Common region filter for all tracks



Tracks

Tracks of the same data type

(eg. different samples of MutationSeq)



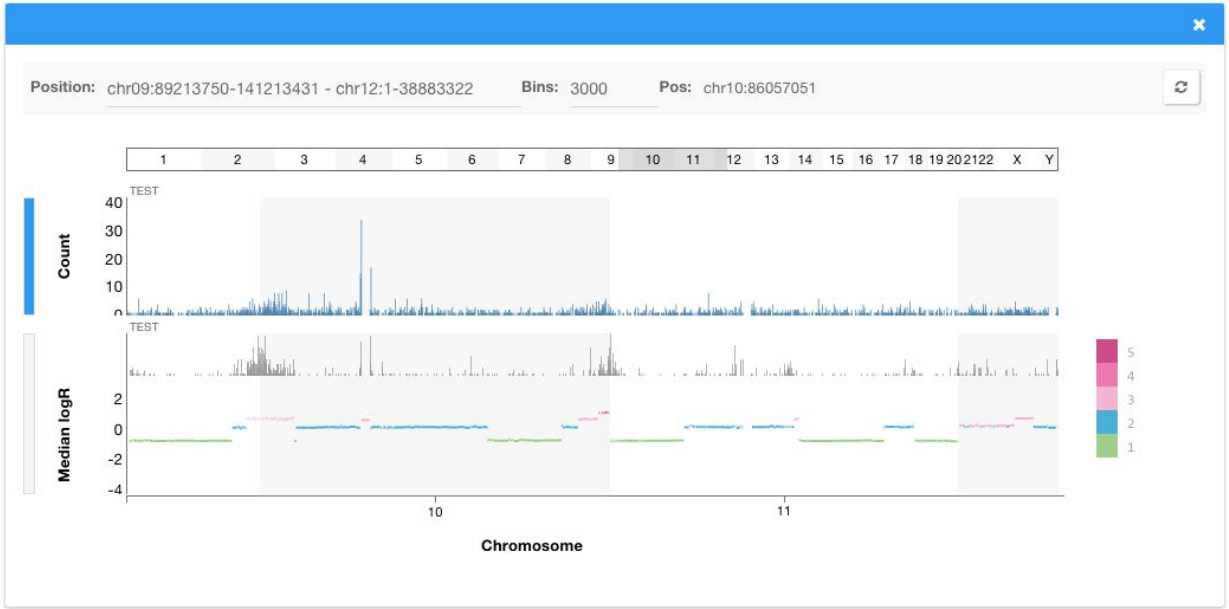
Tracks

Tracks with different data types

(eg. Destruct, MutationSeq, Titan)

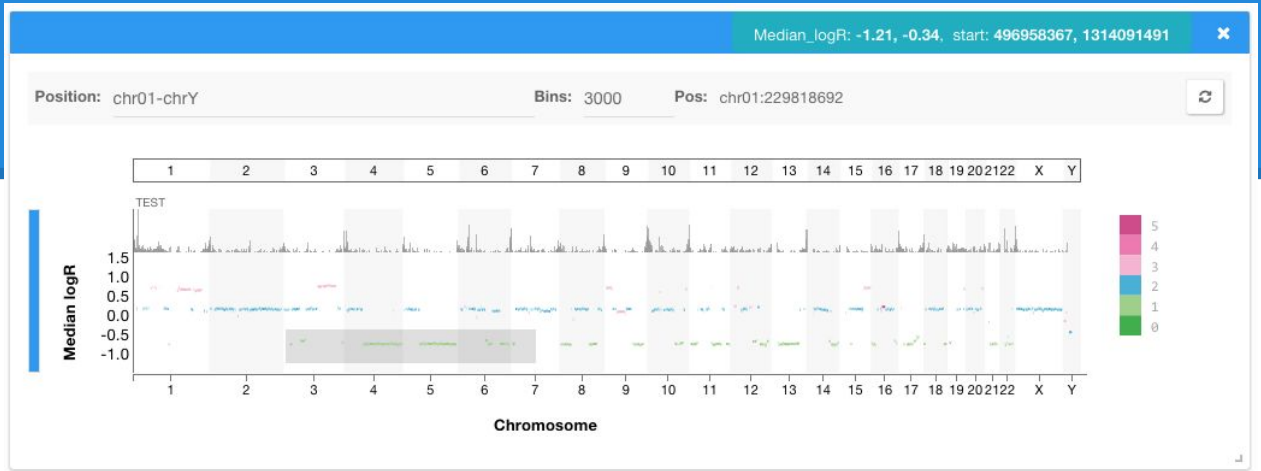
Zooming

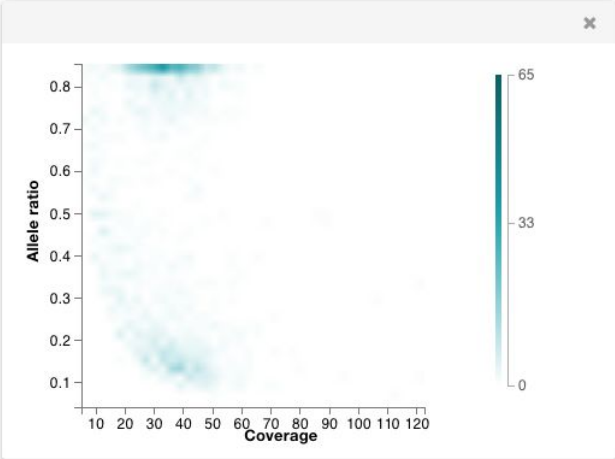
- Zoom in and out using the scroll wheel, or use the genome overview bar to select a specific window to zoom into
- Pan by dragging the selection in genome overview bar



Selection

Select region of genome-wide plot to see corresponding overlapping events in any other plot





Zooming Challenges

Challenges

- Binned data requires executing a query for the given window. We provide the bins, and let Elasticsearch do the counting for us

```
GET denormalized_data/_search
{
  "size": 0,
  "aggs": {
    "chrom01": {
      "terms": {
        "field": "chrom_number",
        "include": "01",
        "size": 10000000
      },
      "aggs": {
        "start": {
          "range": {
            "field": "start",
            "ranges": [
              {
                "from": 0,
                "to": 1029961
              },
              {
                "from": 1029961,
                "to": 2059922
              },
              .....
            ]
          }
        }
      }
    }
    "chrom02": {
      ....
    }
  }
}
```

Zooming Challenges

Challenges

- Binned data requires executing a query for the given window. We provide the bins, and let Elasticsearch do the counting for us


```
"chrom01": {
    "buckets": [
        {
            "key": "01",
            "doc_count": 1859,
            "start": {
                "buckets": [
                    {
                        "key": "0.0-1029961.0",
                        "from": 0,
                        "from_as_string": "0.0",
                        "to": 1029961,
                        "to_as_string": "1029961.0",
                        "doc_count": 5
                    },
                    {
                        "key": "1029961.0-2059922.0",
                        "from": 1029961,
                        "from_as_string": "1029961.0",
                        "to": 2059922,
                        "to_as_string": "2059922.0",
                        "doc_count": 7
                    },
                    {
                        "key": "2059922.0-3089883.0",
                        "from": 2059922,
                        "from_as_string": "2059922.0",
                        "to": 3089883,
                        "to_as_string": "3089883.0",
                        "doc_count": 2
                    },
```

Selection

Existing selection of region in scatterplot

(eg. copy number events that overlap allele ratio 0.73-0.86 and coverage 5-61.27)