Multiscale Visualization of Structural Variants

Armita Safa, Janet Li & Neera Patadia
DNA Sequencing
Genomic Variants

- Sequencing data can be used to find **genomic variants**
- Genomic variant = change in DNA sequence
- Genomic variants can cause **disease**
Pathogenicity

- **Pathogenicity** refers to whether a variant is suspected of causing disease.
- Variant pathogenicity falls on a spectrum.
- Based on evidence and data from empirical and observational studies.
Project Objective

- Thousands upon millions of variants can be identified in single sample

- Interested in developing a tool to visualize genomic structural variant data
Task Abstraction

- Where in the genome did the variant occur?
- Are there diseases/disease risk associated with a given variant?
- What genes are implicated in a given variant?
- Are there other variants that occur in the same region?
- Are there certain areas more prone to having structural variants?
Related Work

● Linear style genome browsers

UCSC Genome Browser

Integrative Genomics Viewer
Related Work

- Multiscale Views
Data and Data Abstraction

Deletion

Insertion

Translocation
Input Datasets

**ClinVar:** Curated database of structural variants with associated pathogenicity classifications
- 150,782 items
- Main attributes:
  - Chromosome (categorical)
  - Position (continuous)
  - Type (categorical)
  - Clinical significance (categorical/ordered)
  - Phenotype list (categorical)
  - Gene list (categorical)

**HG002 variants:** set of high-quality structural variant calls for human individual HG002
- 46,024 items
- Main attributes:
  - Chromosome (categorical)
  - Position (continuous)
  - Type (categorical)
Custom Dataset: Matching Variants

ClinVar

Custom dataset
HG002 SVs with ClinVar matches

HG002 variants
Solution

- Multi-view representation with different levels of details:
  - Circos plot
  - Summary bar charts
  - Linear view
  - Tabular view
  - Interactions to provide details for individual variants

Clinical Significance

- All variants
- Uncertain significance
- Benign
- Likely pathogenic
- Pathogenic
Circos Plot + Linear View
Summary Bar Charts

Clinical Significance
- All variants
- Uncertain significance
- Benign
- Likely pathogenic
- Pathogenic

ClinVar Variants

HG002 Matches
# Match Table

## HG002 Matches

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>Type</th>
<th>Clinical Significance</th>
<th>Similarity</th>
<th>Allele ID</th>
<th>Associated Phenotypes</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19308406</td>
<td>Deletion</td>
<td>Likely pathogenic</td>
<td>9.09</td>
<td>653023</td>
<td>Bare lymphocyte syndrome 2</td>
<td>HGNC:9987</td>
</tr>
<tr>
<td>1</td>
<td>102511070</td>
<td>Deletion</td>
<td>Benign</td>
<td>4.55</td>
<td>1212109</td>
<td>-</td>
<td>HGNC:2961</td>
</tr>
<tr>
<td>1</td>
<td>102511060</td>
<td>Deletion</td>
<td>Benign</td>
<td>4.76</td>
<td>1212109</td>
<td>-</td>
<td>HGNC:2961</td>
</tr>
<tr>
<td>1</td>
<td>102455150</td>
<td>Duplication</td>
<td>Uncertain significance</td>
<td>7.69</td>
<td>642080</td>
<td>Charcot-Marie-Tooth disease, axonal, type 2O</td>
<td>HGNC:2961</td>
</tr>
<tr>
<td>1</td>
<td>100857299</td>
<td>Deletion</td>
<td>Benign</td>
<td>100.0</td>
<td>1241732</td>
<td>-</td>
<td>HGNC:2898</td>
</tr>
<tr>
<td>1</td>
<td>95570219</td>
<td>Duplication</td>
<td>Pathogenic</td>
<td>8.0</td>
<td>921345</td>
<td>DICER1-related pleuropulmonary blastoma cancer predisposition syndrome</td>
<td>HGNC:17098</td>
</tr>
<tr>
<td>1</td>
<td>92679004</td>
<td>Duplication</td>
<td>Uncertain significance</td>
<td>6.15</td>
<td>198351</td>
<td>ANKRD1-related dilated cardiomyopathy; Cardiovascular phenotype; Primary dilated cardiomyopathy</td>
<td>HGNC:15819</td>
</tr>
<tr>
<td>1</td>
<td>105803315</td>
<td>Deletion</td>
<td>Likely pathogenic</td>
<td>9.09</td>
<td>205161</td>
<td>Junctional epidermolysis bullosa, non-Herlitz type</td>
<td>HGNC:2184</td>
</tr>
<tr>
<td>1</td>
<td>10616</td>
<td>copy number loss</td>
<td>Likely pathogenic</td>
<td>0.0</td>
<td>435724</td>
<td>-</td>
<td>HGNC:329</td>
</tr>
<tr>
<td>1</td>
<td>977156</td>
<td>Deletion</td>
<td>Benign</td>
<td>100.0</td>
<td>666917</td>
<td>-</td>
<td>HGNC:329</td>
</tr>
</tbody>
</table>
Implementation

Pre-processing

- Filter ClinVar dataset
- **Match** HG002 events to ClinVar variants
  - Same chromosome
  - Distance < 20
  - Similarity score

```
ACTTGTCTTTATGC
ACT___G___TTATA___C
```
Implementation

Visualization

React JS

GOSLING
Grammar Of Scalable Linked Interactive Nucleotide Graphics

DB
Limitations & Future Work

● Click events not supported by Gosling.js
  ○ Select specific variants and present details in table
● Custom glyphs for different variant types
  ○ Current solutions are not ideal
● Match variants from user input
Conclusion

- We have created a multi-scale visualization tool for examining the clinical relevance of SVs
- Created a custom dataset + new derived attribute for annotating SVs
- SV data is shown on multiple scales
- Interactive features allow users to explore data at different levels of detail
References


