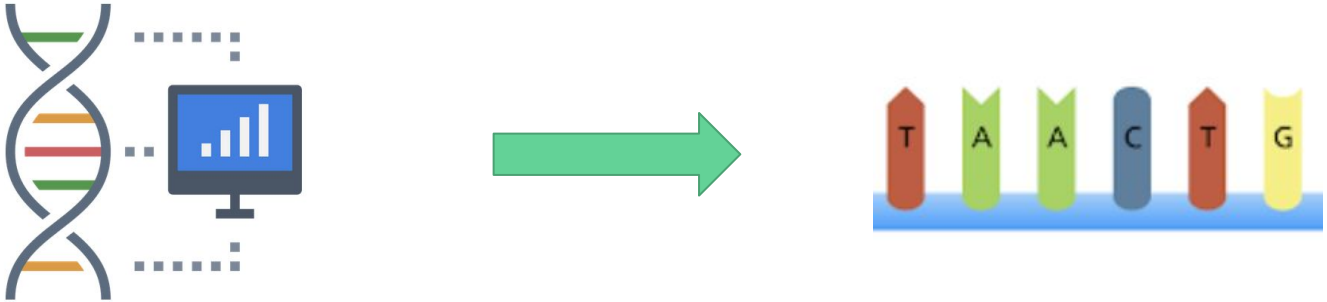
Multiscale Visualization of Structural Variants

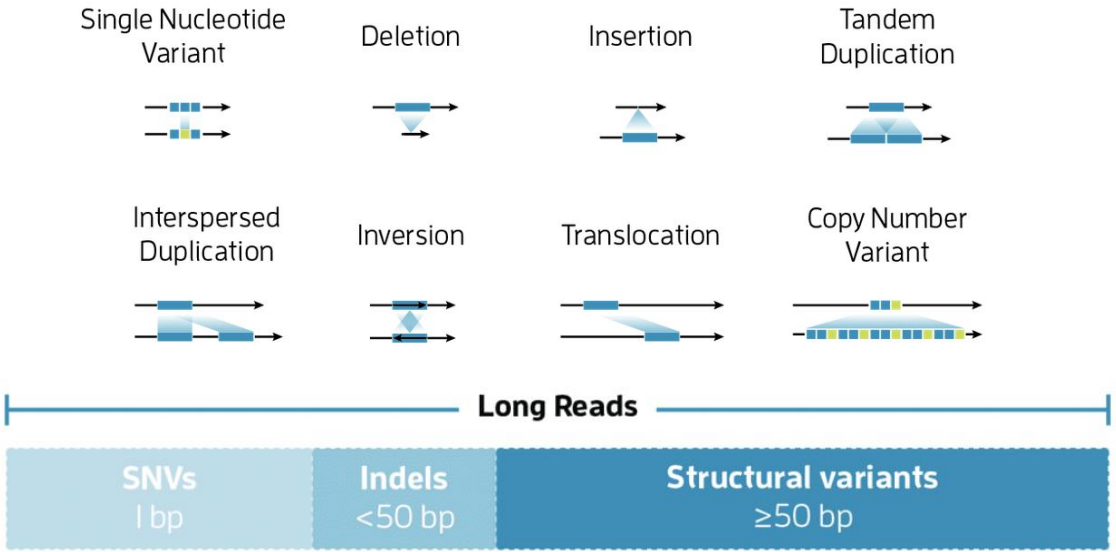
Armita Safa, Janet Li & Neera Patadia

DNA Sequencing



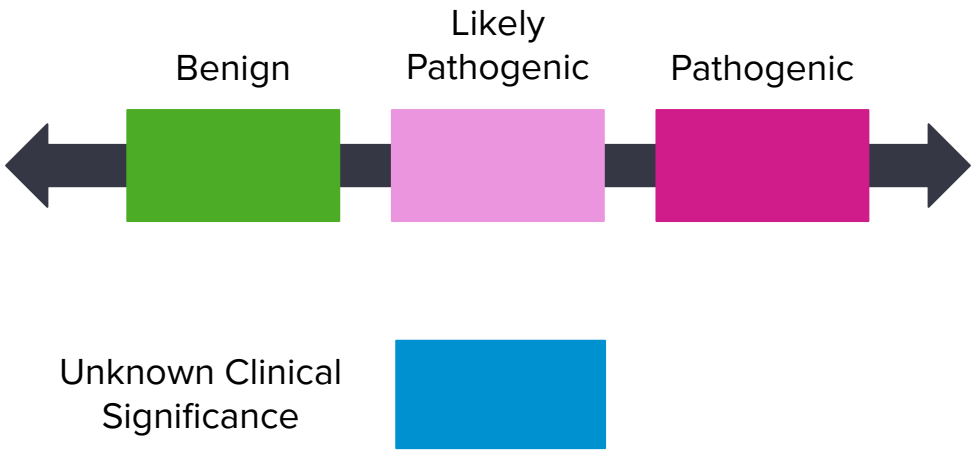
Genomic Variants

- 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

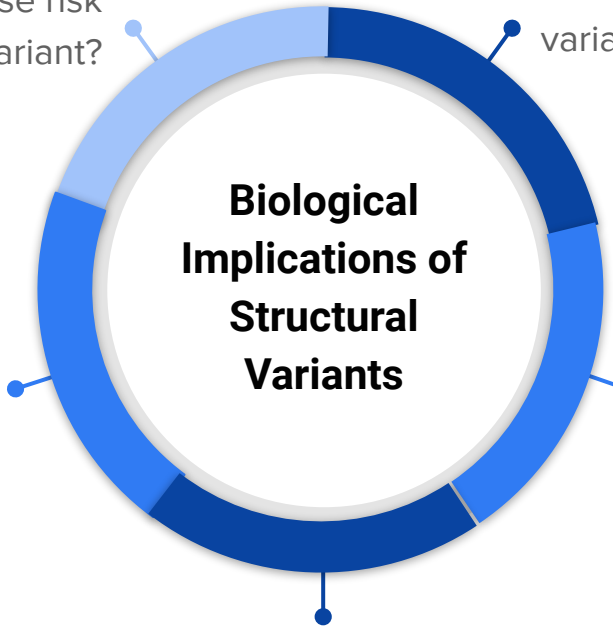
- Thousands upon millions of variants can be identified in single sample

```
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF     ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G       A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:
20     17330   .         T       A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A       G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T       .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC     G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

CHROM	POS	END	Similarity	AlleleID	Type	HGNC_ID	ClinicalSignificance	PhenotypeList
1	19308406	19308426	9.09	653023	Deletion	HGNC:9987	Likely pathogenic	Bare lymphocyte syndrome 2
1	102511070	102511112	4.55	1212109	Deletion	HGNC:2961	Benign	-
1	102511060	102511100	4.76	1212109	Deletion	HGNC:2961	Benign	-
1	102455150	102455199	7.69	642080	Duplication	HGNC:2961	Uncertain significance	Charcot-Marie-Tooth disease, axonal, type 20
1	100657299	100657331	100.0	1241732	Deletion	HGNC:2698	Benign	-
1	95570219	95570241	8.0	921345	Duplication	HGNC:17098	Pathogenic	DICER1-related pleuropulmonary blastoma cancer predisposition syndrome
1	92679004	92679066	6.15	198351	Duplication	HGNC:15819	Uncertain significance	ANKRD1-related dilated cardiomyopathy;Cardiovascular

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

Task Abstraction



Are there diseases/disease risk associated with a given variant?

Where in the genome did the variant occur?

What genes are implicated in a given variant?

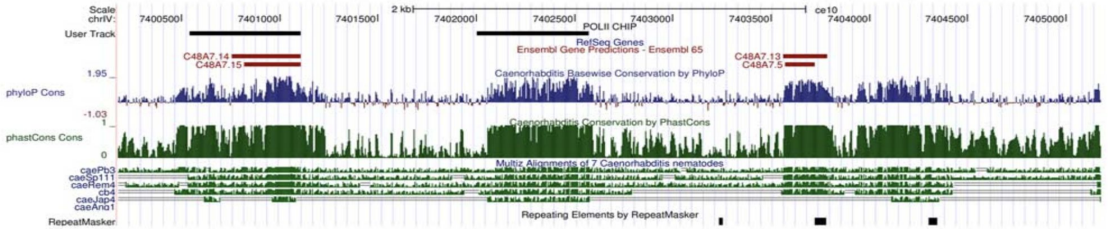
Are there other variants that occur in the same region?

Are there certain areas more prone to having structural variants?

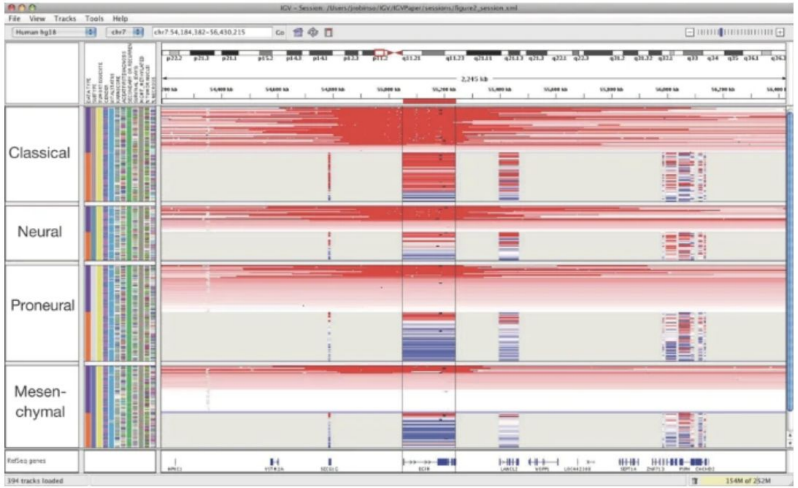
Related Work

- Linear style genome browsers

UCSC Genome Browser

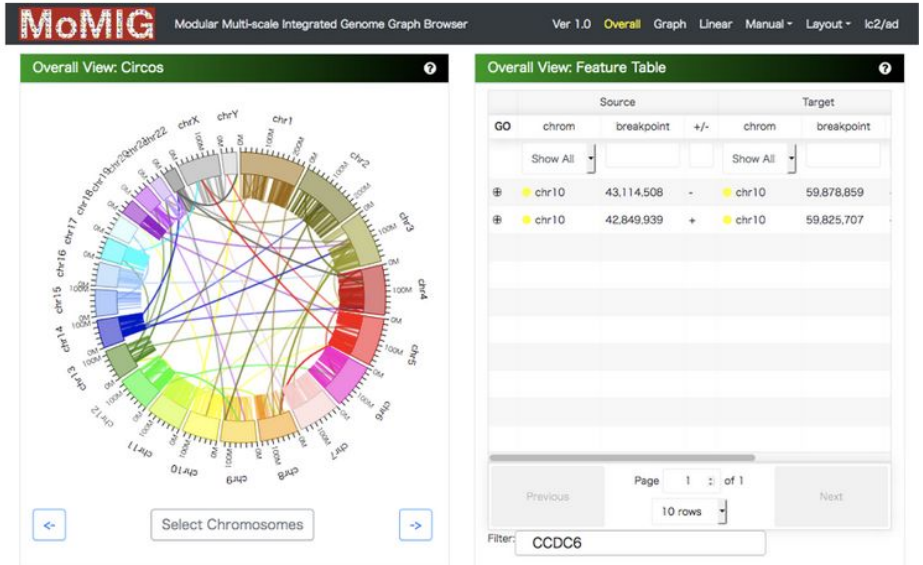


Integrative Genomics Viewer

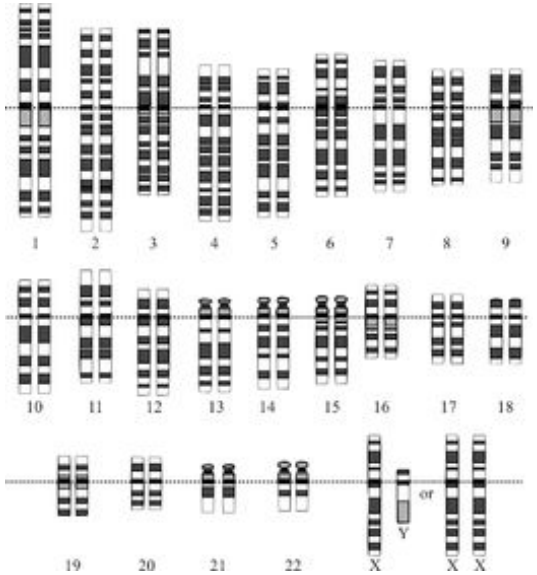


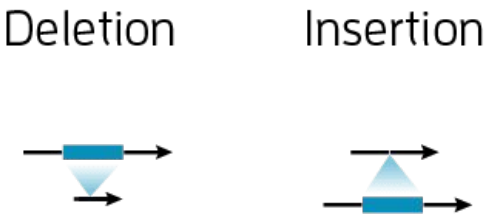
Related Work

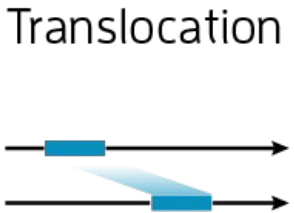
- Multiscale Views



Data and Data Abstraction







Input Datasets

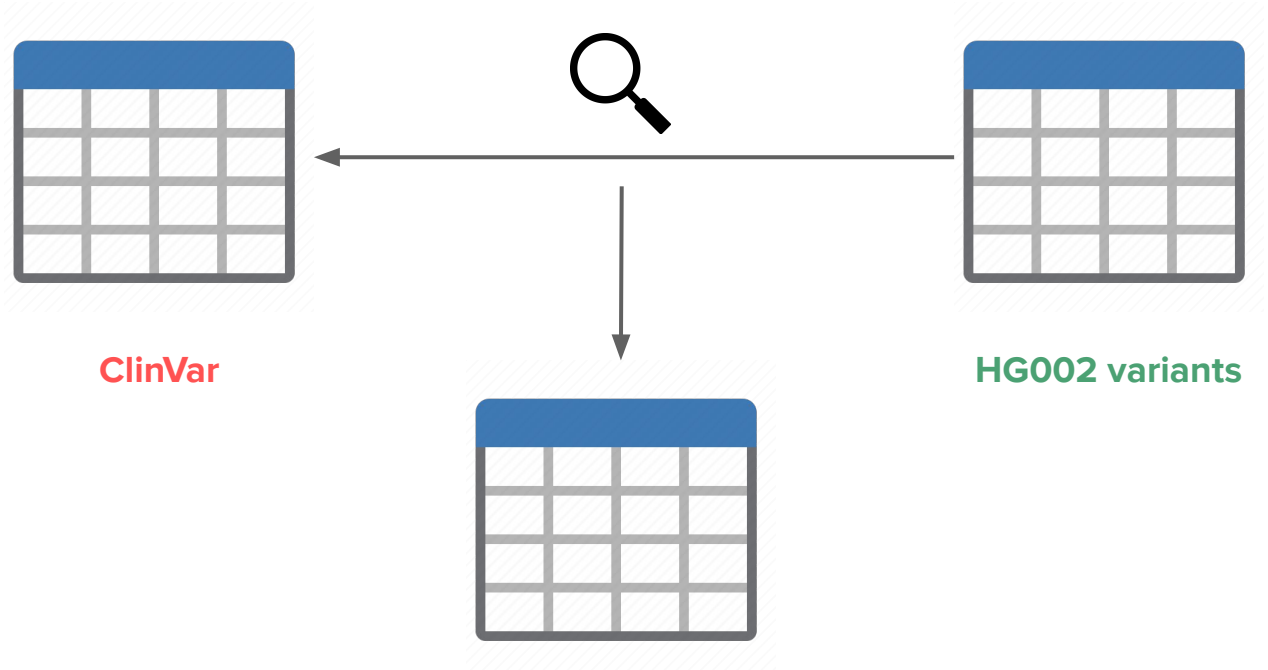
ClinVar: Curated database of structural variants with associated pathogenicity classifications

- 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)

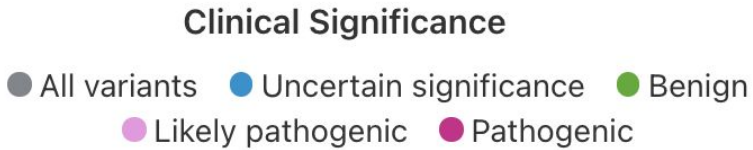
Custom Dataset: Matching Variants



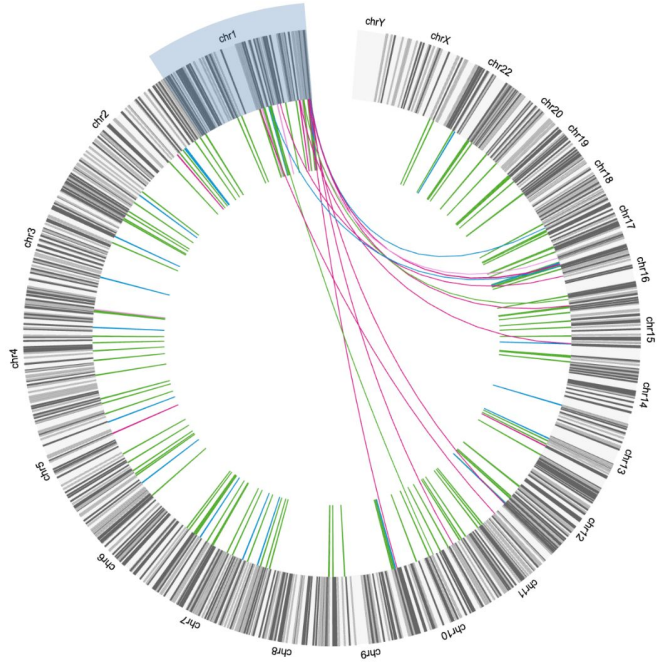
Custom dataset
HG002 SVs with ClinVar matches

Solution

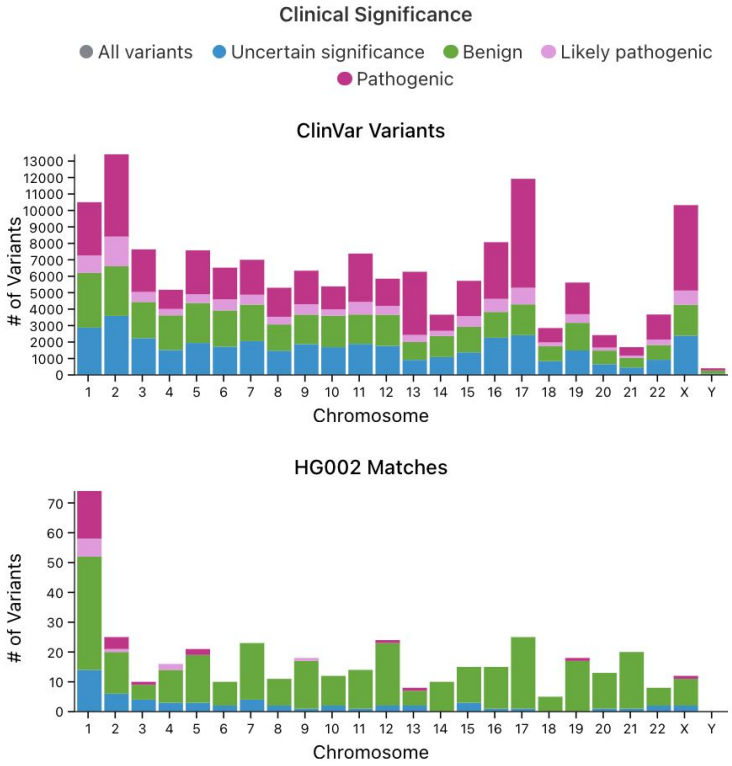
- Multi-view representation with different levels of details:
  - Circos plot
  - Summary bar charts
  - Linear view
  - Tabular view
  - Interactions to provide details for individual variants



Circos Plot + Linear View



Summary Bar Charts



Match Table

HG002 Matches

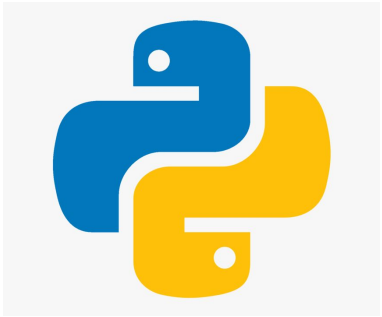
Chr	Position	Type	Clinical Significance	Similarity	Allele ID	Associated Phenotypes	Gene
1	19308406	Deletion	Likely pathogenic	9.09	653023	Bare lymphocyte syndrome 2	HGNC:9987
1	102511070	Deletion	Benign	4.55	1212109	-	HGNC:2961
1	102511060	Deletion	Benign	4.76	1212109	-	HGNC:2961
1	102455150	Duplication	Uncertain significance	7.69	642080	Charcot-Marie-Tooth disease, axonal, type 2O	HGNC:2961
1	100657299	Deletion	Benign	100.0	1241732	-	HGNC:2698
1	95570219	Duplication	Pathogenic	8.0	921345	DICER1-related pleuropulmonary blastoma cancer predisposition syndrome	HGNC:17098
1	92679004	Duplication	Uncertain significance	6.15	198351	ANKRD1-related dilated cardiomyopathy;Cardiovascular phenotype;Primary dilated cardiomyopathy	HGNC:15819
1	105803315	Deletion	Likely pathogenic	9.09	205161	Junctional epidermolysis bullosa, non-Herlitz type	HGNC:2194
1	10616	copy number loss	Likely pathogenic	0.0	435724	-	-
1	977156	Deletion	Benign	100.0	656917	-	HGNC:329

<< < > >> Page **1 of 8** | Go to page: 1 Show 10

Implementation

Pre-processing

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

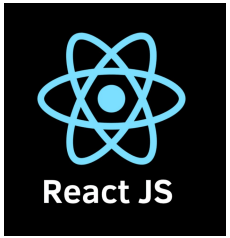


A C T T G T C T T A T G C

A C T _ G _ _ T T A _ _ C

Implementation

Visualization



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

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 Data-Driven Documents. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

Karolchik, D., Hinrichs, A. S., & Kent, W. J. (2009). The UCSC Genome Browser. *Current Protocols in Bioinformatics*, *28*(1), 1.4.1-1.4.26. https://doi.org/10.1002/0471250953.bi0104s28

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42(Database issue), D980–D985. https://doi.org/10.1093/nar/gkt1113

L'Yi, S., Wang, Q., Lekschas, F., & Gehlenborg, N. (2021). Gosling: A Grammar-based Toolkit for Scalable and Interactive Genomics Data Visualization. OSF Preprints. https://doi.org/10.31219/osf.io/6evmb

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. Nature Biotechnology, 29(1), 24–26. https://doi.org/10.1038/nbt.1754

Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y., & Kasahara, M. (2019). MoMI-G: Modular multi-scale integrated genome graph browser. BMC Bioinformatics, 20(1), 548. https://doi.org/10.1186/s12859-019-3145-2