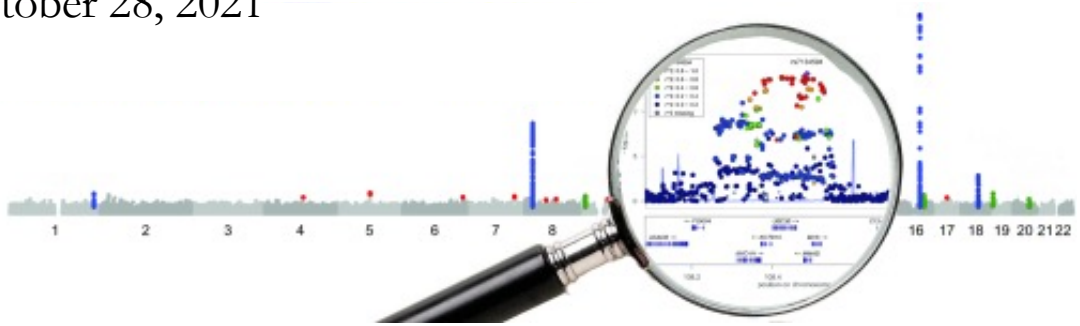
LocusZoom.js:

Interactive and embeddable visualization of genetic association study results

Ryan Welch and Andy Boughton

University of Michigan Center for Statistical Genetics

October 28, 2021



Acknowledgements

Michigan
Daniel Taliun
Peter VandeHaar
Alan Kwong
Chris Clark
Matthew Flickinger
Sarah Gagliano Taliun
Josh Weinstock
Laura Scott
Gonçalo Abecasis
Mike Boehnke
Hyun Min Kang

HuGeAmp Team
Ben Alexander
Jeffrey Massung
Kenneth Bruskiewicz
Marc Duby
Maria Costanzo
Noel Burtt
Jason Flannick

CMDGA
Parul Kudtarkar
Kyle Gaulton

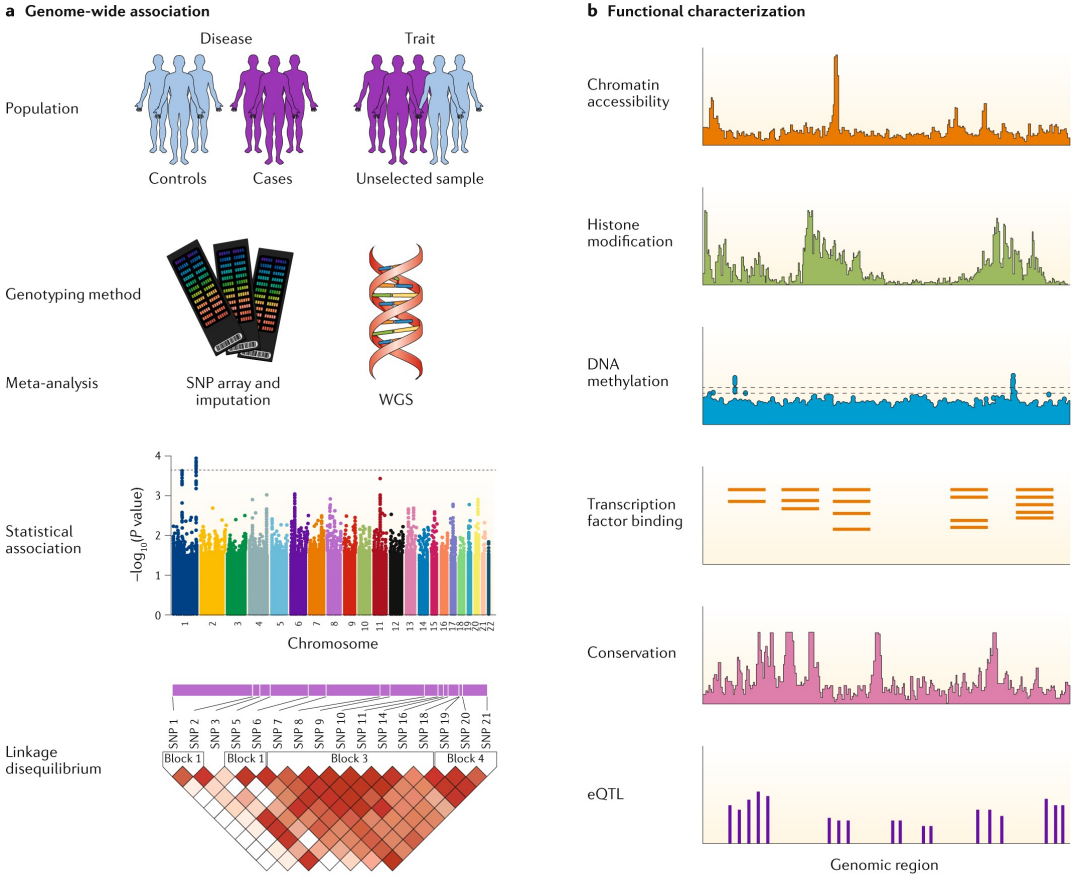
Others
Karen Mohlke
John Kemp

HuGeAMP

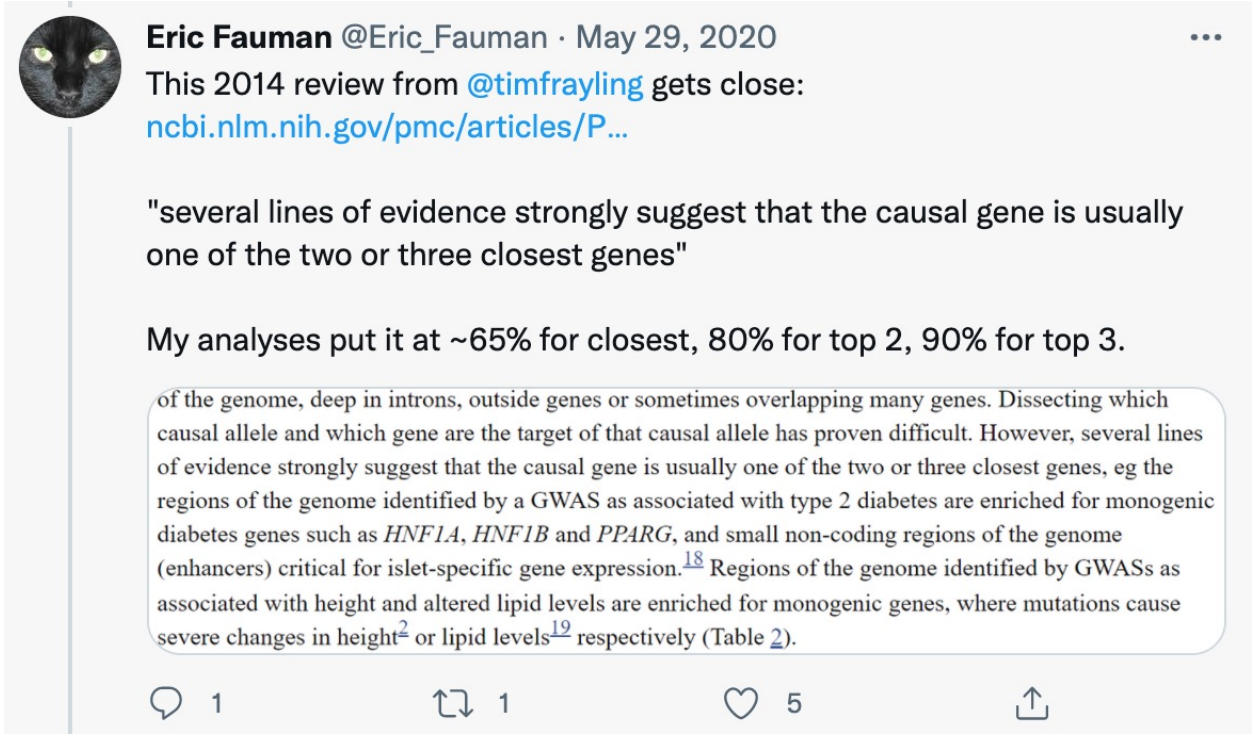
Common Metabolic Diseases Genome Atlas

GWAS: From correlation to causality

- Genome-Wide Association Studies find variants correlated with disease...
  - ...but the molecular mechanism is not always clear
- Results must be understood in biological context



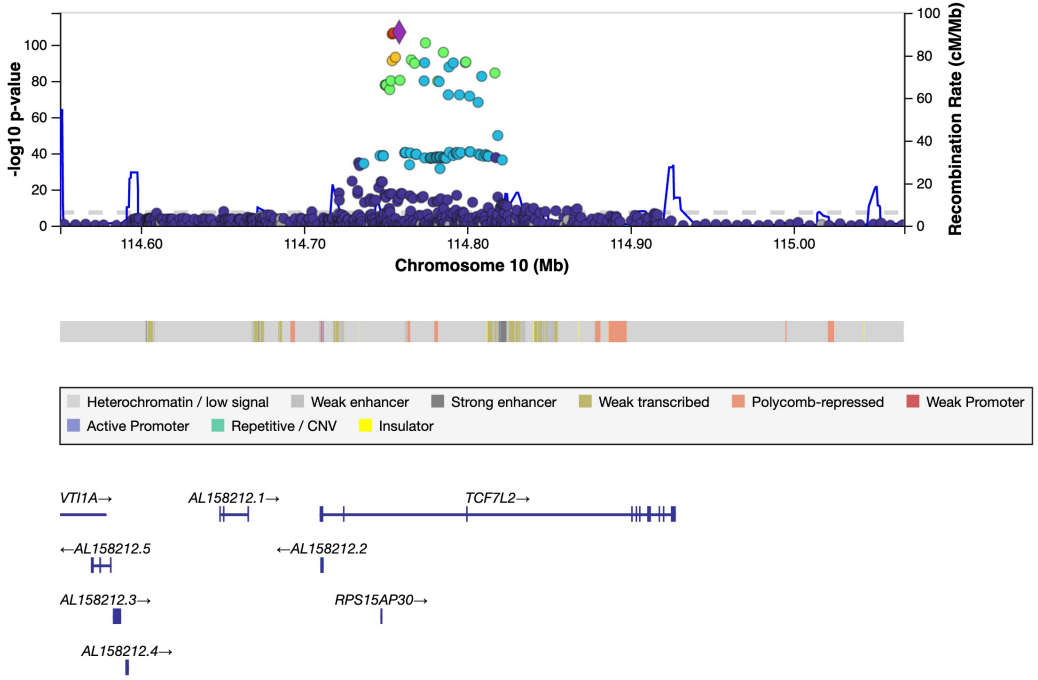
https://www.ebi.ac.uk/gwas/
https://doi.org/10.1093/nar/gky1120
https://doi.org/10.1038/s41576-019-0127-1



https://twitter.com/Eric_Fauman/status/1266315707249242117

LocusZoom: History and Purpose

- Designed to visualize GWAS summary statistics in a region
- A command line version has been widely used for ~10 years
- New interactive LocusZoom.js supports exploration and web-based data sharing

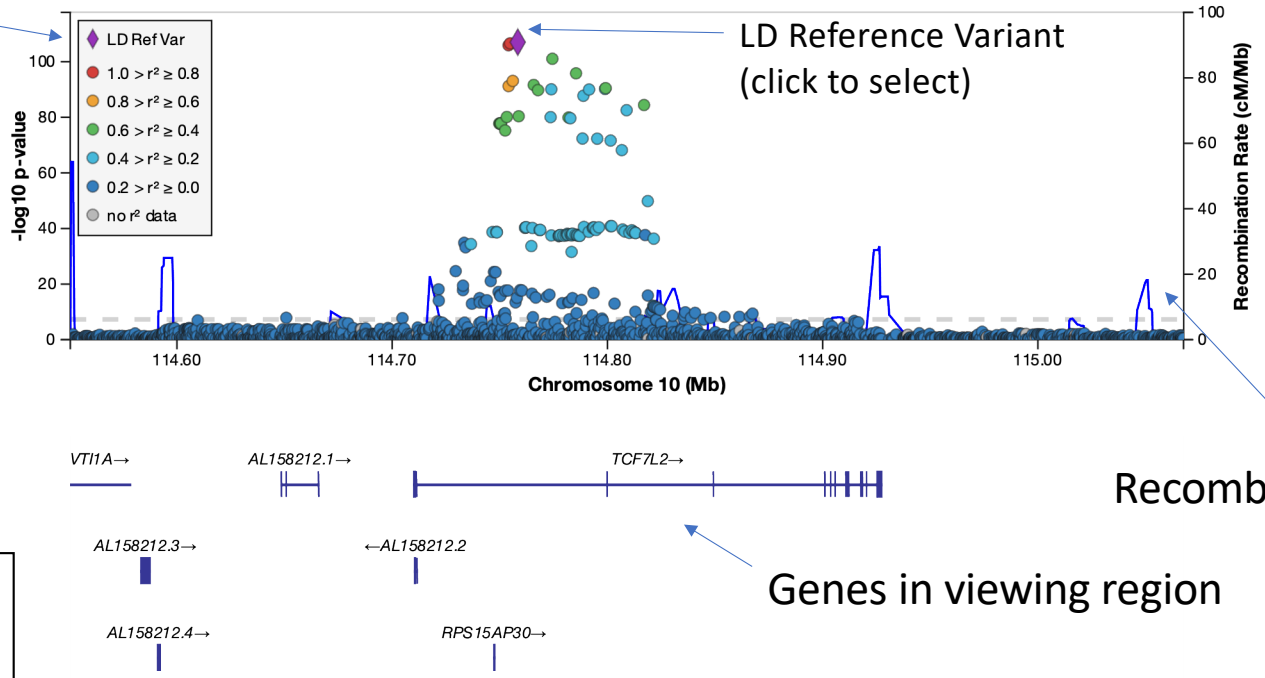


https://doi.org/10.1093/bioinformatics/btab186

https://github.com/statgen/locuszoom

How to read a LocusZoom Plot

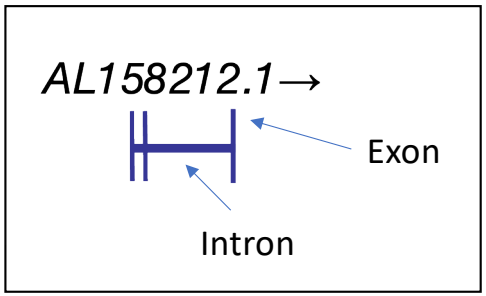
Color by Linkage Disequilibrium (LD)



LD Reference Variant (click to select)

Recombination rate

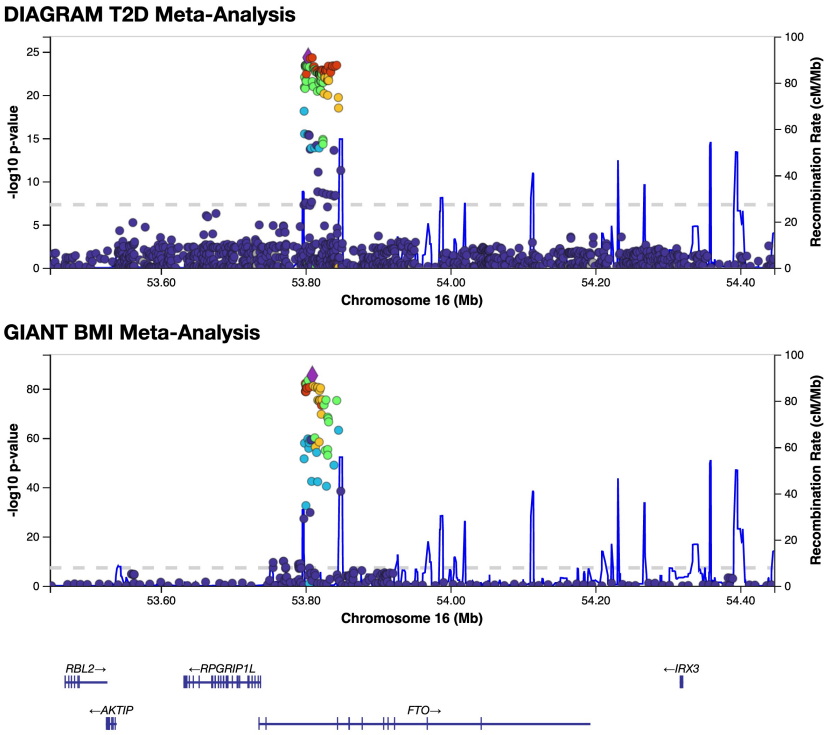
Genes in viewing region



Key Features

Designed for data sharing

- Re-usable genetic visualization widget- not just tied to one website
- Allows comparing different studies and kinds of data
- Explore many regions of a study- not limited to static premade images
- Highly customizable control of point display, interactivity, etc

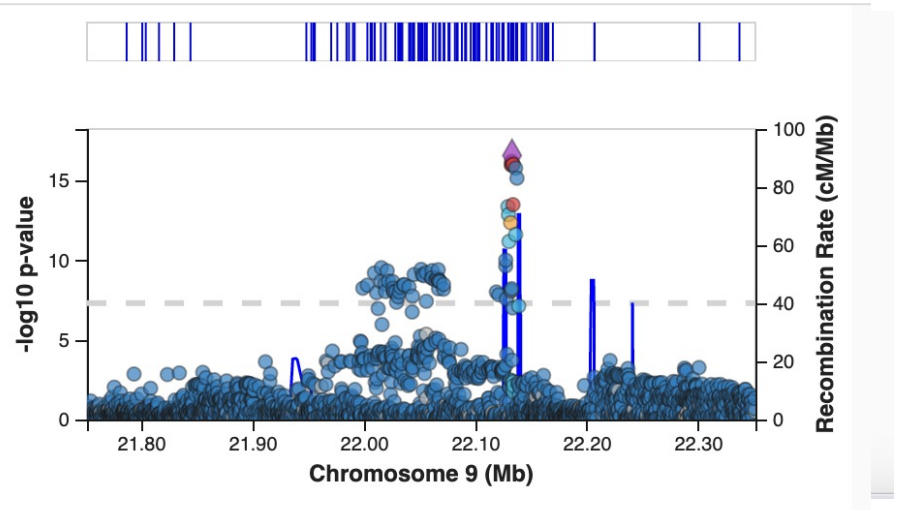


PheWeb T2DKP FIVEx

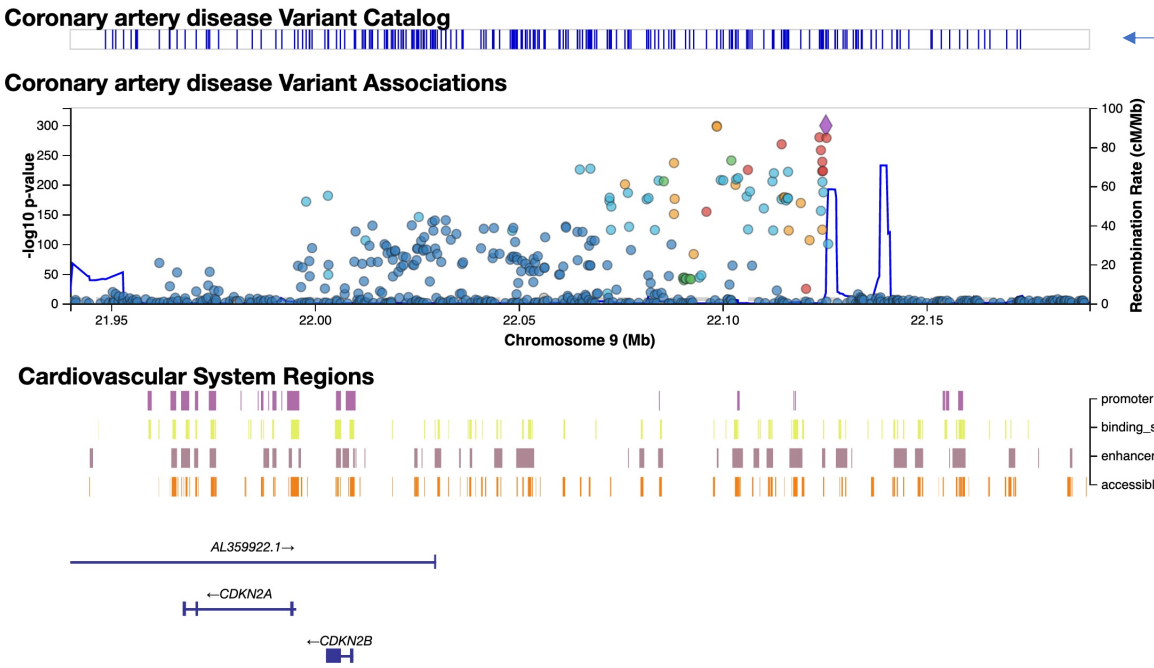
my.locuszoom.org LocalZoom

Showing the best annotations available

- Downloadable tools often end up using outdated annotation information
- LocusZoom.js fetches updated annotation data as it becomes available
  - Genes, recomb rate, GWAS catalog
  - Support for GRCh37 and GRCh38
- As datasets grow, we continue to evaluate usability



Add optional annotation tracks

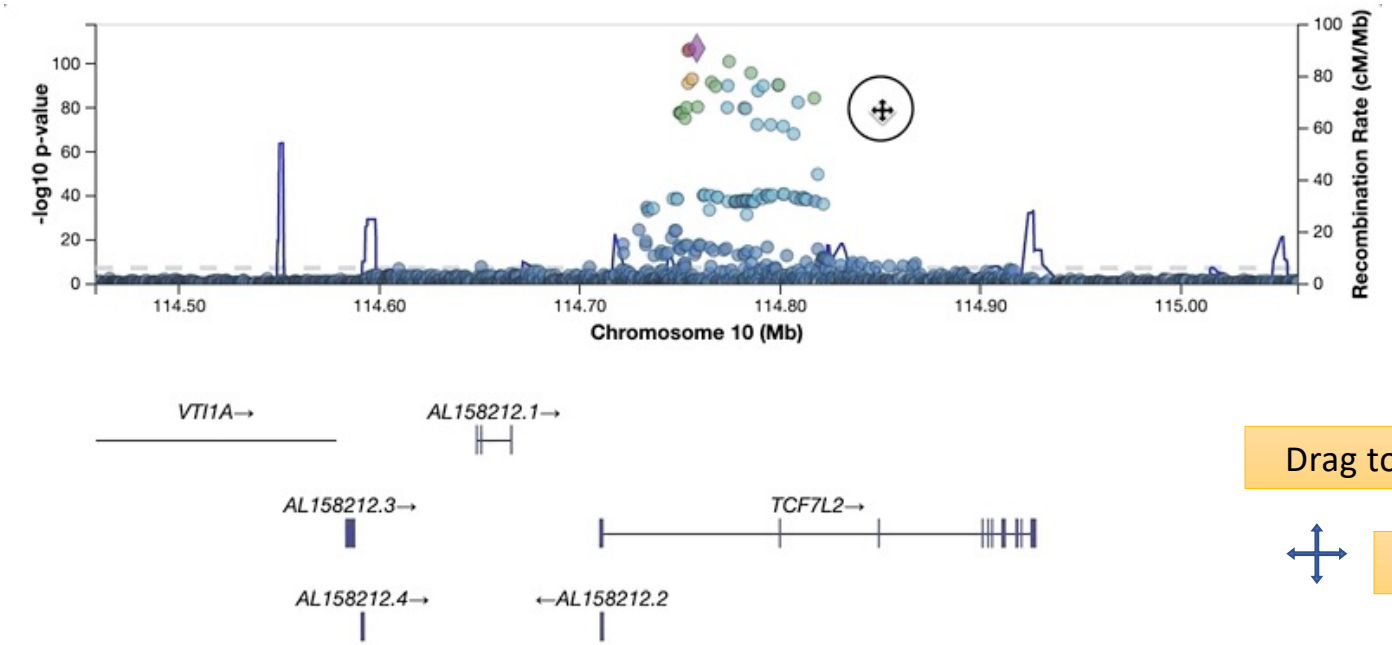


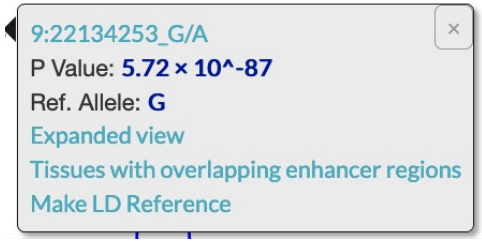
Highlights variants for which there are significant claims in the EBI GWAS catalog

Show multiple BED-track style annotations for chromatin state, chromatin accessibility, or enhancer/promoter interactions

https://t2d.hugeamp.org/region.html?chr=8&end=118238953&phenotype=T2D&start=117912512

Interactively control the view



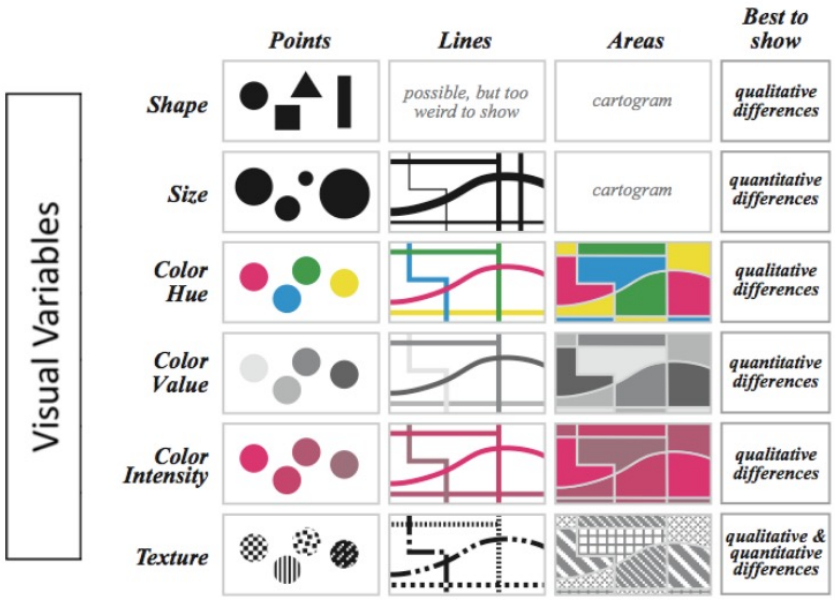


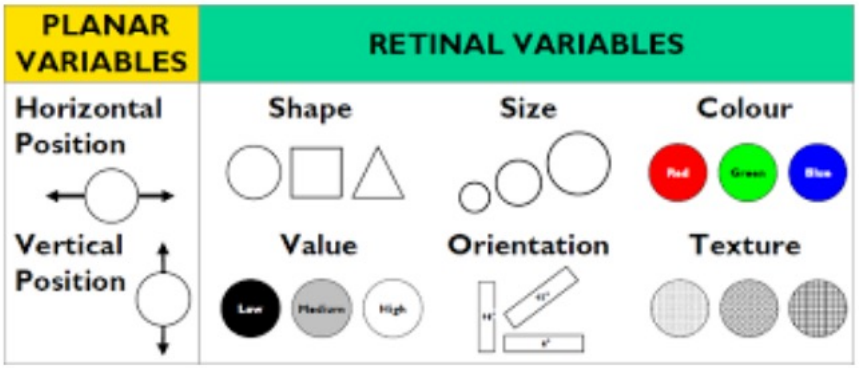
Drag to pan

Scroll to zoom

Click to interact

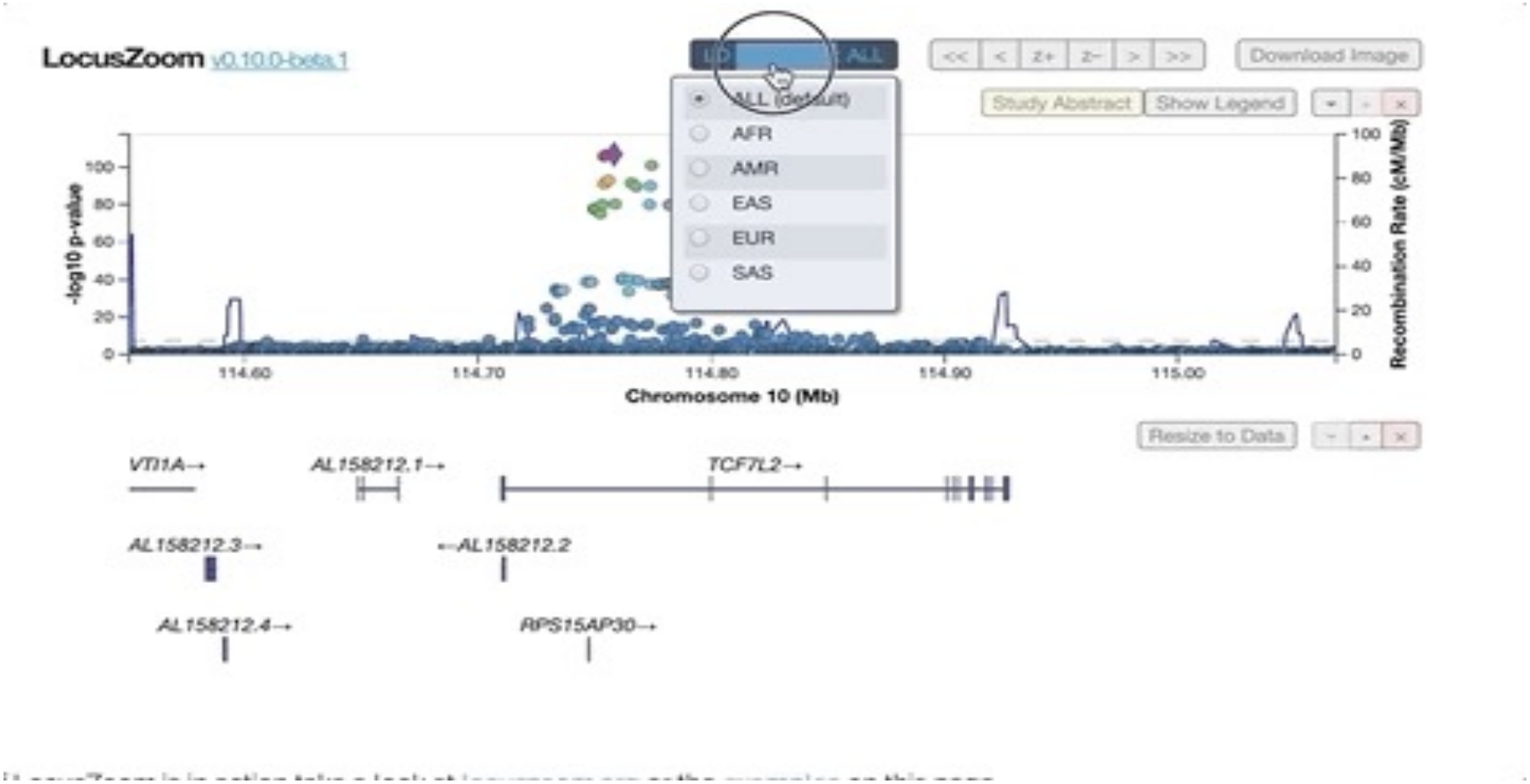
Interactivity lets us show more information





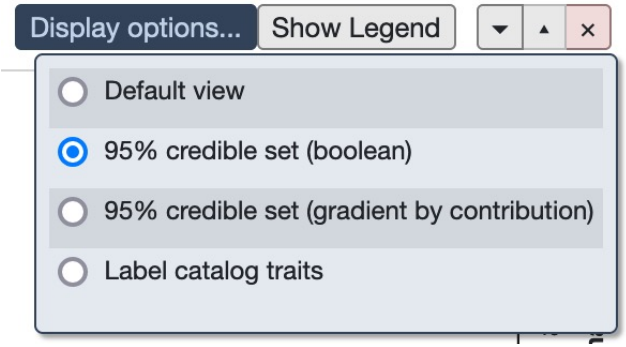
http://euclid.psych.yorku.ca/www/psy6135/lectures/Language.pdf

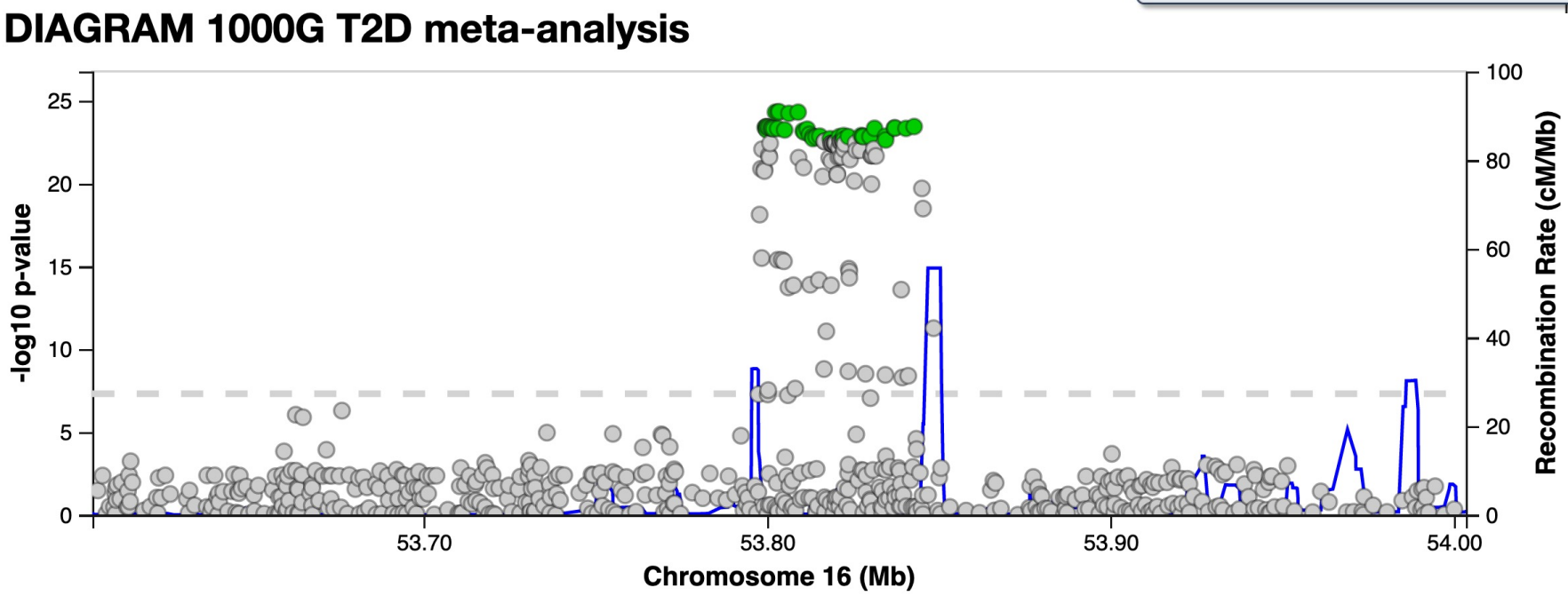
Choose population-specific LD



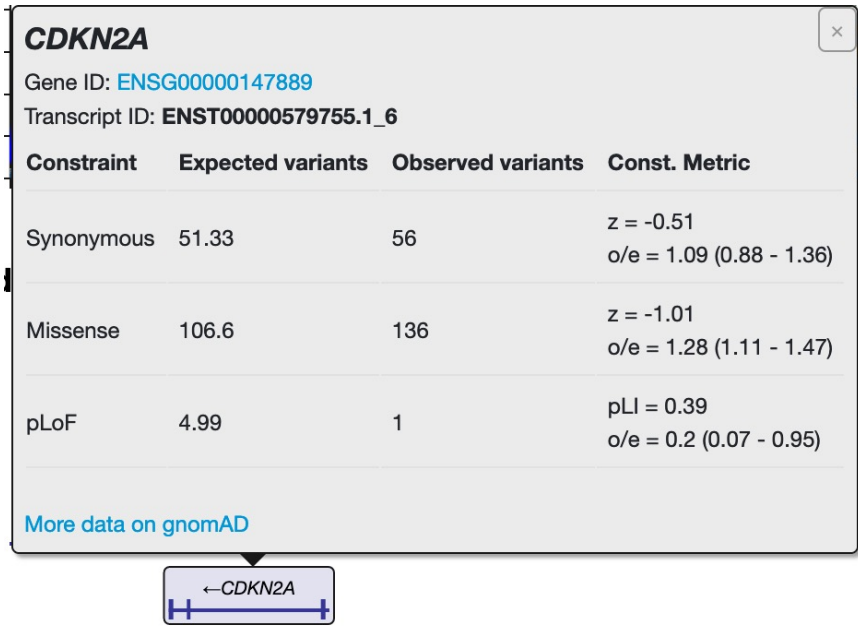
Powered by Michigan LDServer and the 1000 Genomes Project

Overlay different annotations



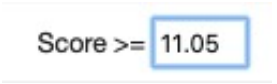


Tooltips provide deeper context

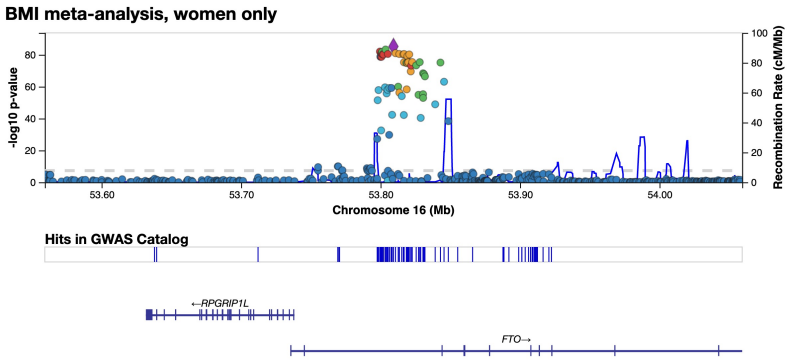


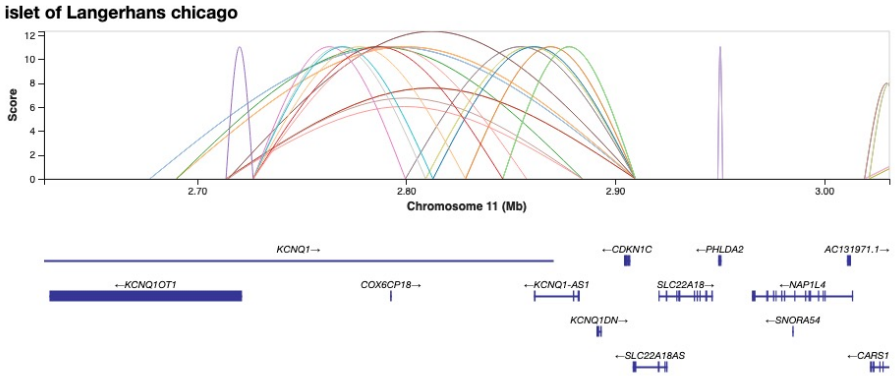
- Click or mouse over any element to produce a tooltip
- Allows showing metrics, quality metrics, or outbound links to other databases
- Action buttons can select a variant and update calculations (like LD Reference variant)

Guiding User Attention



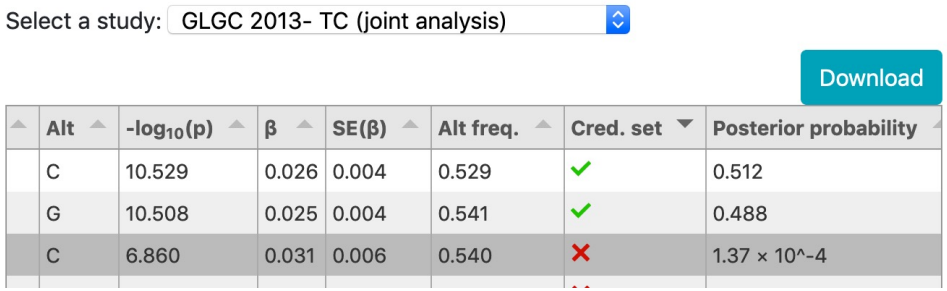
- Interactive "filter" and "match" features make it possible to show/hide tracks, or interactively draw connections between two datasets
- Helps users dig through multivariate datasets to find items of interest





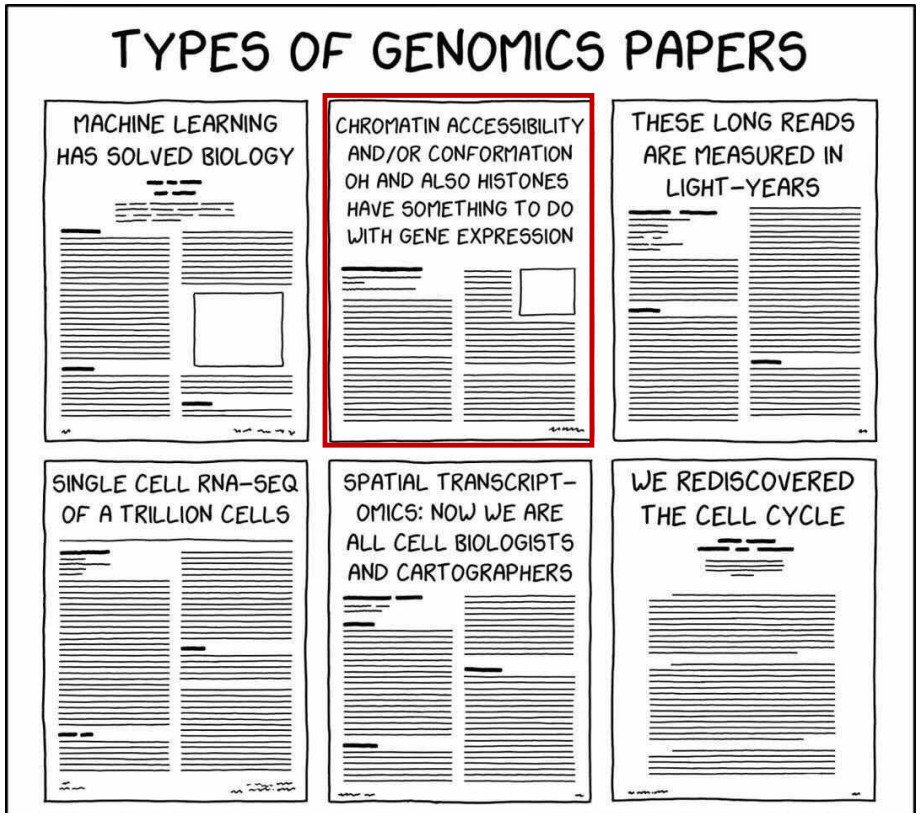
Connect to other representations

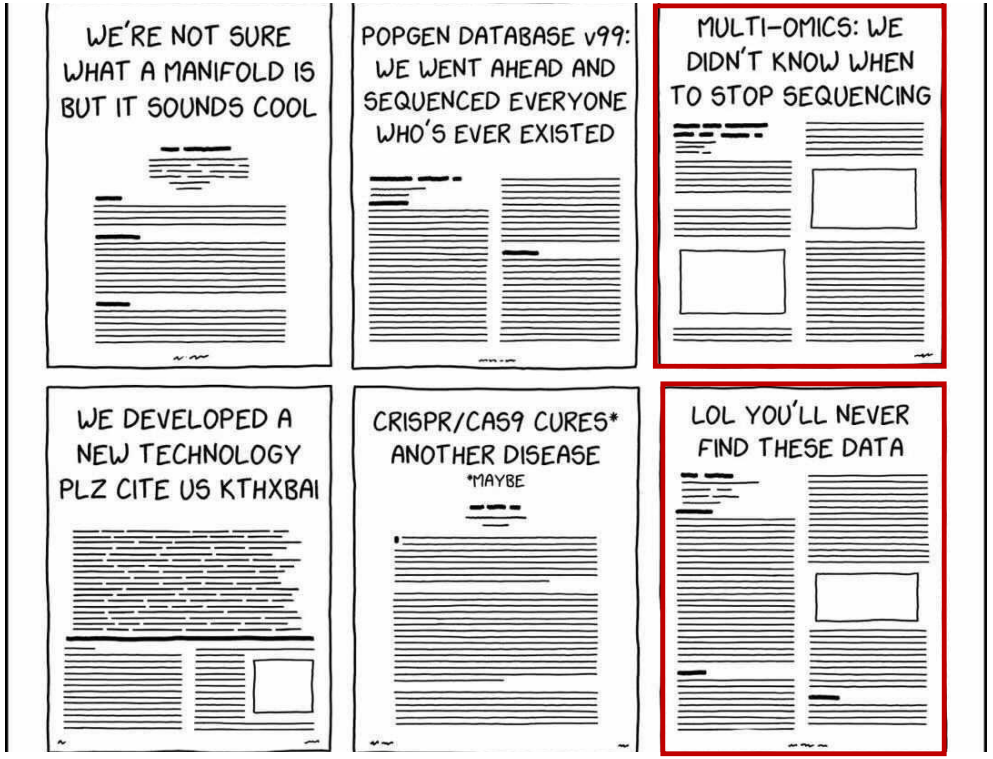
- LocusZoom.js can talk to other visualizations on the same web page, enabling rich comparisons that update as the view changes
- Interact with LocusZoom, drag plot region, and instantly see results update in a table
- Additional results can be computed in browser, then saved as a file for further analysis
- Plots can be saved and shared as high-resolution images



Applications

Visualizing diverse datatypes across the field



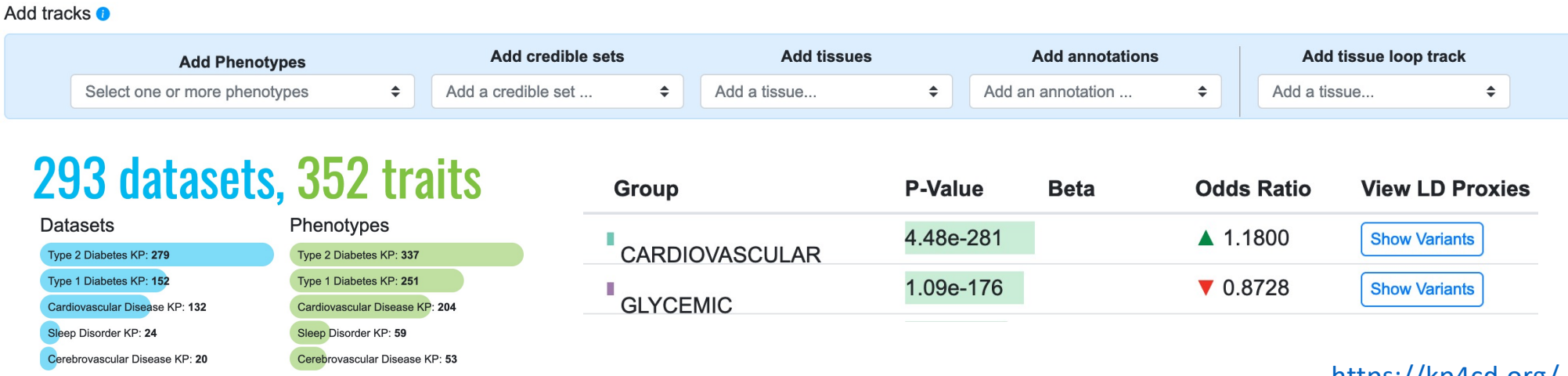


- The field is evolving fast
- Must allow people to make connections across existing and novel datasets, for many kinds of data

https://twitter.com/iddux/status/1389268071911669769

CMDKP/ HuGeAMP: Knowledge Portal Network

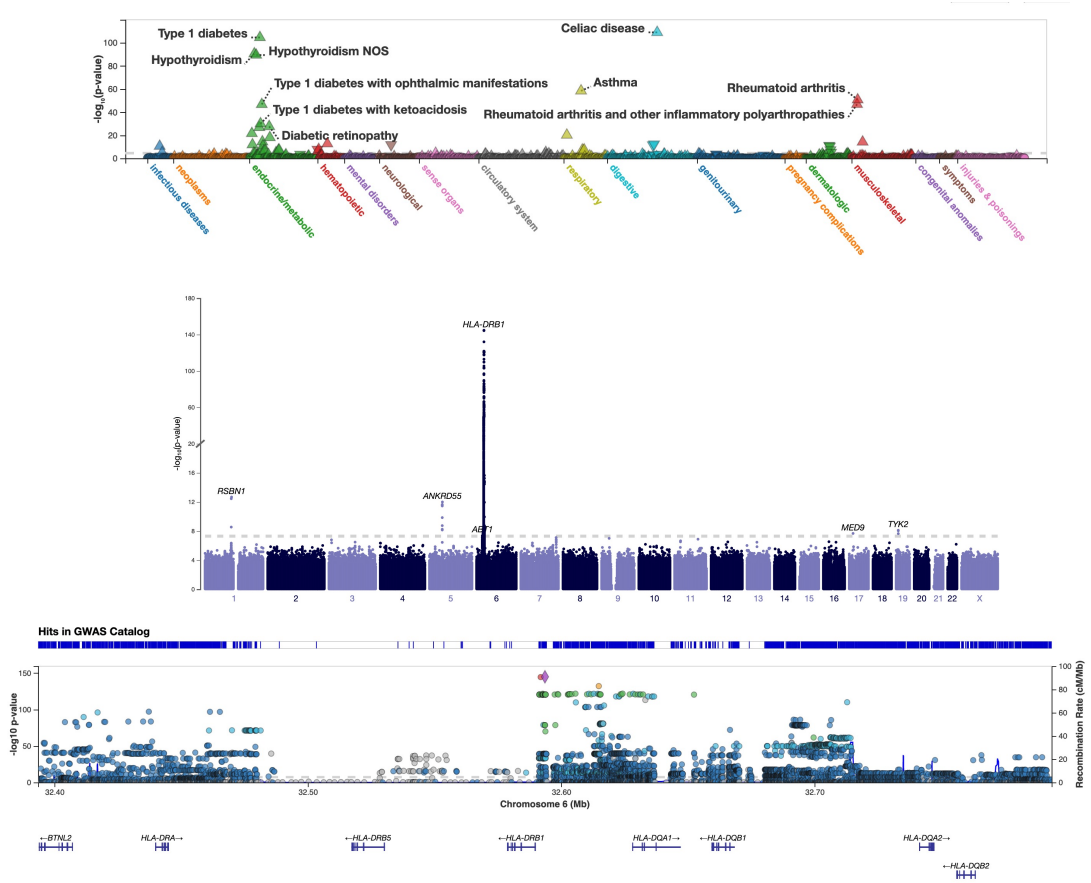
- Aggregator of many datasets, phenotypes, and annotations
- Users can choose from a wide range of data and visualization types to add to the plot based on their phenotype and tissue of interest
- Plots synchronize with data tables to provide summary and detail views



https://kp4cd.org/

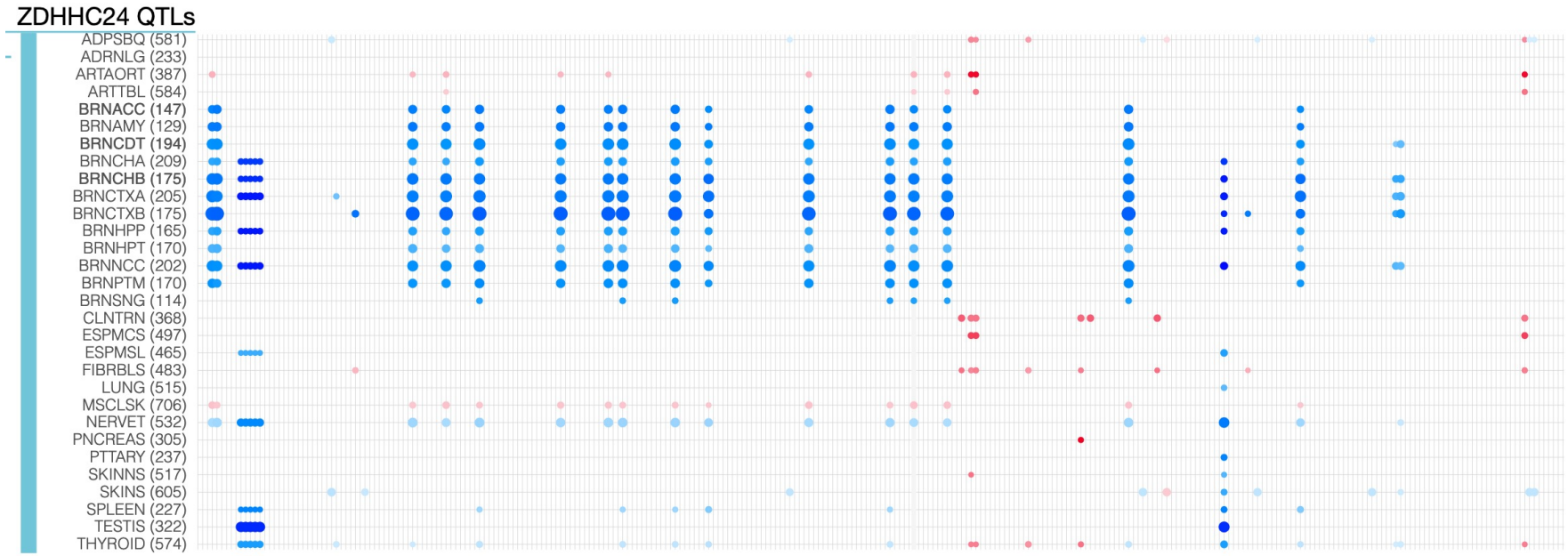
PheWeb: Phenome Wide Associations

- Connect associations across many phenotypes
- Quickly jump between all-studies view (PheWAS), single study (Manhattan plot), and region view (LocusZoom)
- A popular tool in the era of biobank studies (Finngen, UKBB, etc)



https://pheweb.org
https://doi.org/10.1038/s41588-020-0622-5

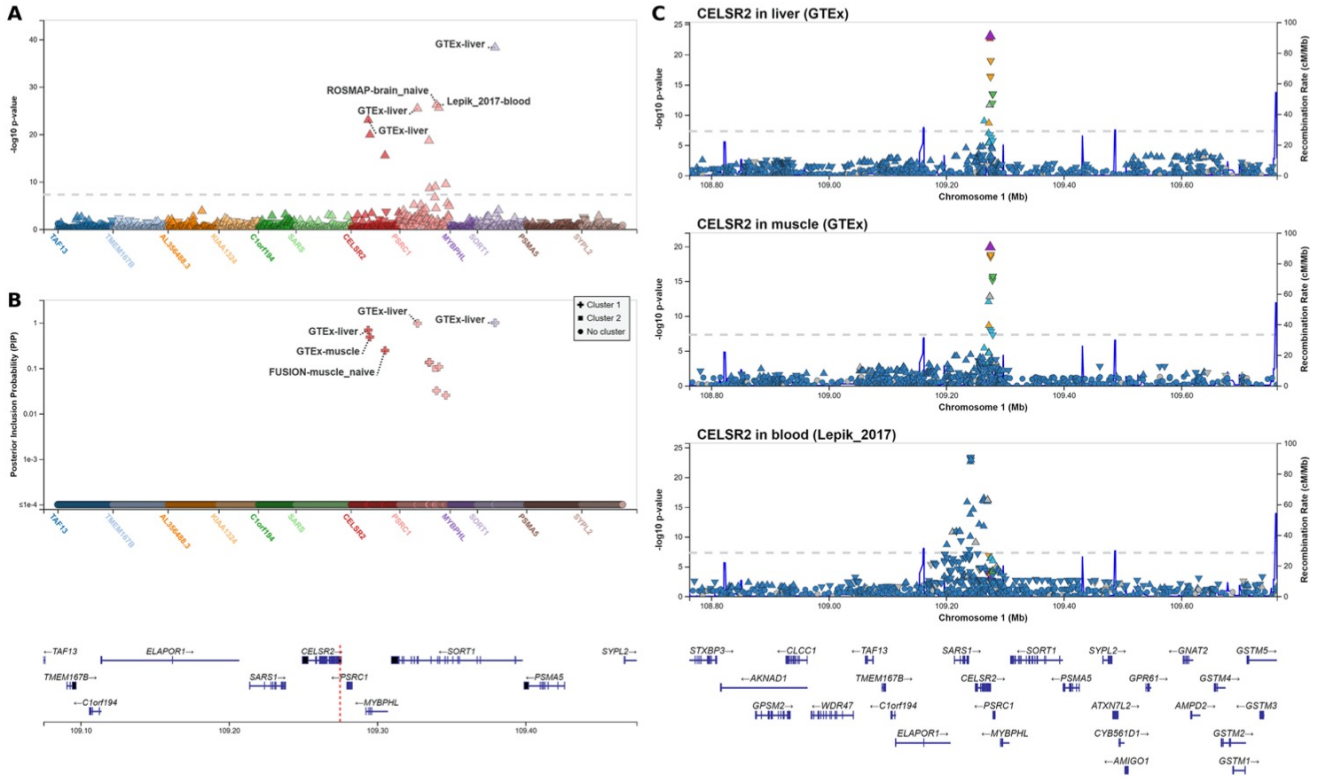
eQTLs: Often presented in terms of variants...



https://gtexportal.org/home/locusBrowserVCPage/chr11_66545555_A_G_b38

FIVEx: (e)QTL browser

- Draw comparisons for same variant across different tissues
- Is a strong QTL signal an artifact of other variants in high LD?
  - Use credible sets to narrow down unique signals
- Interactively switch between 15+ eQTL studies and dozens of possible views

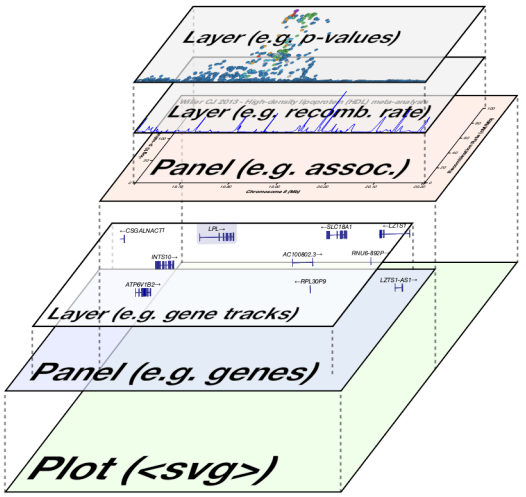


https://fivex.sph.umich.edu/
https://doi.org/10.1093/bioinformatics/btab614

Implementation and usage

LocusZoom.js: Modular and reusable

- LocusZoom.js can be embedded in your web page, and can draw data from a wide range of sources
- It is based on standard web technologies that work with many sites
- No special tooling is required, but it will work with advanced tools if you use them
- Our standard REST APIs provide common data (like genes) for builds GRCh37 and GRCh38



https://github.com/statgen/locuszoom/

Adding LocusZoom.js to a web page

```
// First, specify how to find and load data
var apiBase = "https://portaldev.sph.umich.edu/api/v1/";
var data_sources = new LocusZoom.DataSources()
  .add("assoc", ["AssociationLZ", {url: apiBase + "statistic/single/", params: { source: 45, id_field: "variant" }}]);
  ...

// Second, specify how this information will be displayed. (premade config options)
var layout = LocusZoom.Layouts.get("plot", "standard_association", { state: { genome_build: 'GRCh37' } });

// Last, draw the plot in a specified location (div container) as defined in the HTML
window.plot = LocusZoom.populate("#lz-plot", data_sources, layout);
```

https://portaldev.sph.umich.edu/docs/api/v1/#overview-of-api-endpoints
https://github.com/statgen/locuszoom/blob/develop/README.md
https://github.com/statgen/locuszoom/blob/develop/docs/simplest_template.html

Visualizations require data!

- LocusZoom.js can read data from many sources, but the data needs to be given to the plot in a certain expected format
- We recommend storing data in a way that supports region queries:
  - Tabix files are useful for indexing a single study (can be queried directly via S3)
  - REST API server for querying one out of many studies ("get me data for phenotype x") or protecting sensitive data

https://portaldev.sph.umich.edu/docs/api/v1/#overview-of-api-endpoints
https://statgen.github.io/locuszoom/docs/guides/data_retrieval.html
https://statgen.github.io/locuszoom/docs/api/module-ext_lz-tabix-source.html

LocusZoom APIs

- LocusZoom.js requires various pieces of data to function
  - Association statistics (variants & -$\log_{10}$ p-values)
  - Gene definitions (start/end of exons, introns)
  - Recombination rates
  - LD between lead variant and all other variants
- To get these data, LocusZoom.js queries our public API servers
  - LocusZoom API server: genes, recombination rates, GWAS catalog
  - LDServer: LD between variants (based on 1000G reference panel)
- Both API servers expose HTTP REST APIs
  - LocusZoom API: https://portaldev.sph.umich.edu/docs/api/v1/#overview-of-api-endpoints
  - LDServer: https://portaldev.sph.umich.edu/playground

https://github.com/statgen/LDServer
https://github.com/statgen/locuszoom-api

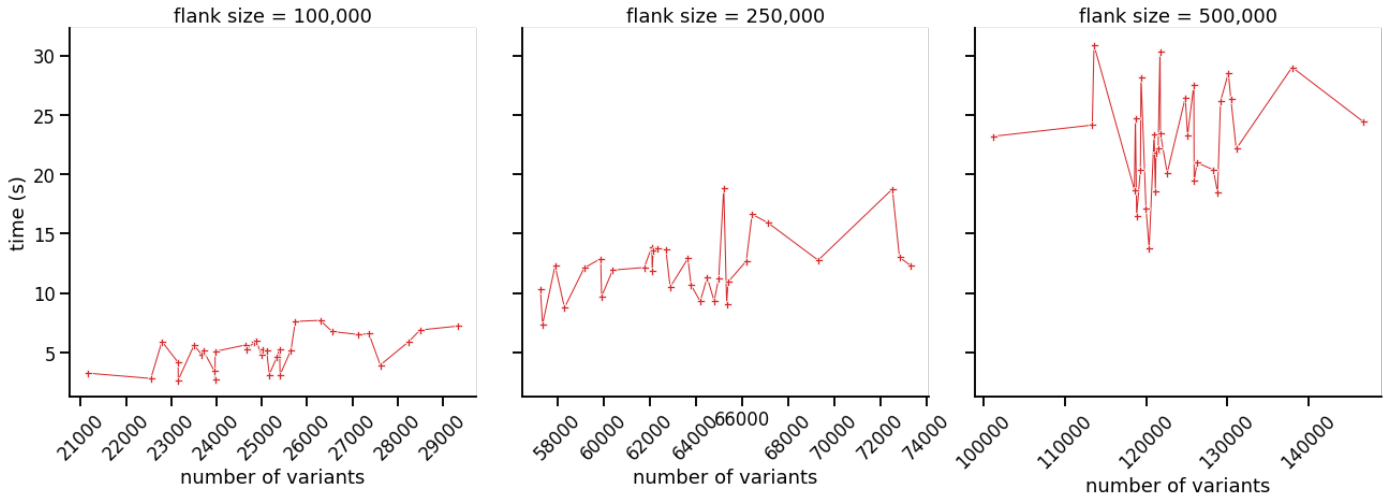
LDServer: high performance LD queries

- Retrieving LD quickly is important for *interactive* visualization
  - At 1000G reference panel size: relatively fast
  - At TOPMed size: much slower, 100-200k+ samples
- Multiple strategies used to speed up serving LD:
  - Redis (in-memory LRU cache) to store commonly requested LD, saving computation time
  - Savvy[1] (SAV) genotype files: columnar storage, high compression, supports region queries
  - Fast approximate method for calculating LD (emeraLD[2])
  - OpenMP for parallel calculation of LD once genotypes retrieved from disk
  - Calculations performed in C++ using armadillo + OpenBLAS (or MKL)
  - Multiple workers to field requests in parallel; increase worker count to support higher load

[1] https://github.com/statgen/savvy
[2] Quick C, Fuchsberger C, Taliun D, Abecasis G, Boehnke M, Kang HM. emeraLD: rapid linkage disequilibrium estimation with massive datasets. Bioinformatics. 2019;35: 164–166.

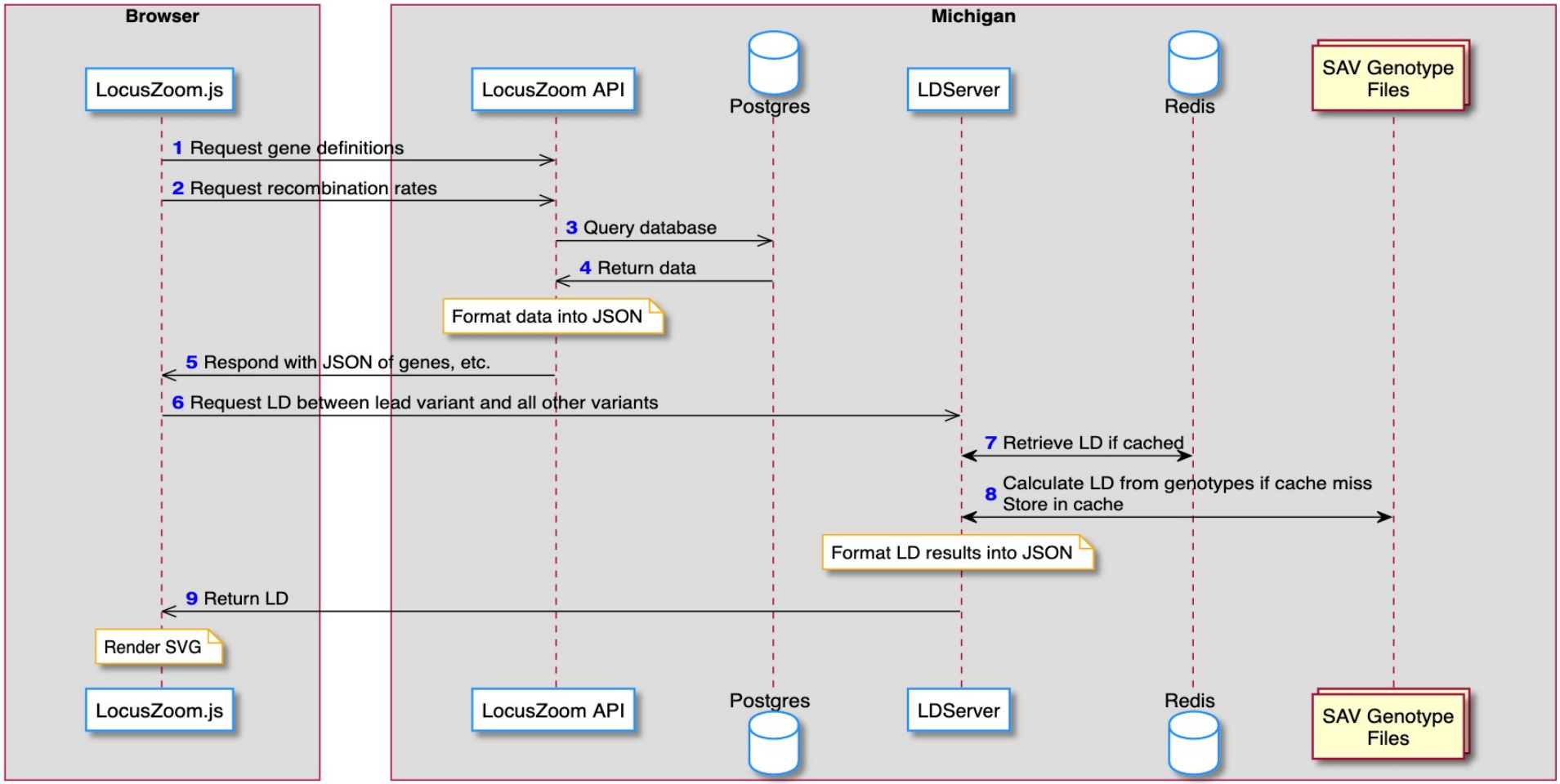
TOPMed LD benchmarking

- Using TOPMed freeze 8 (200k+ samples!), no singletons/doubletons, randomly sampled regions, single thread



- For smaller regions: query performance important
- For large regions: caching becomes critical

How does it all work together?



Exploratory analysis with LocusZoom

In-browser analysis enabled by summary statistics

- Many analyses are possible at a genetic locus with only basic summary statistics
  - Score statistic for each single variant
  - Covariance matrix of score statistics

$$S = G^T(y - \hat{\mu}_0)$$

$$\Psi = G^T \hat{P} G$$

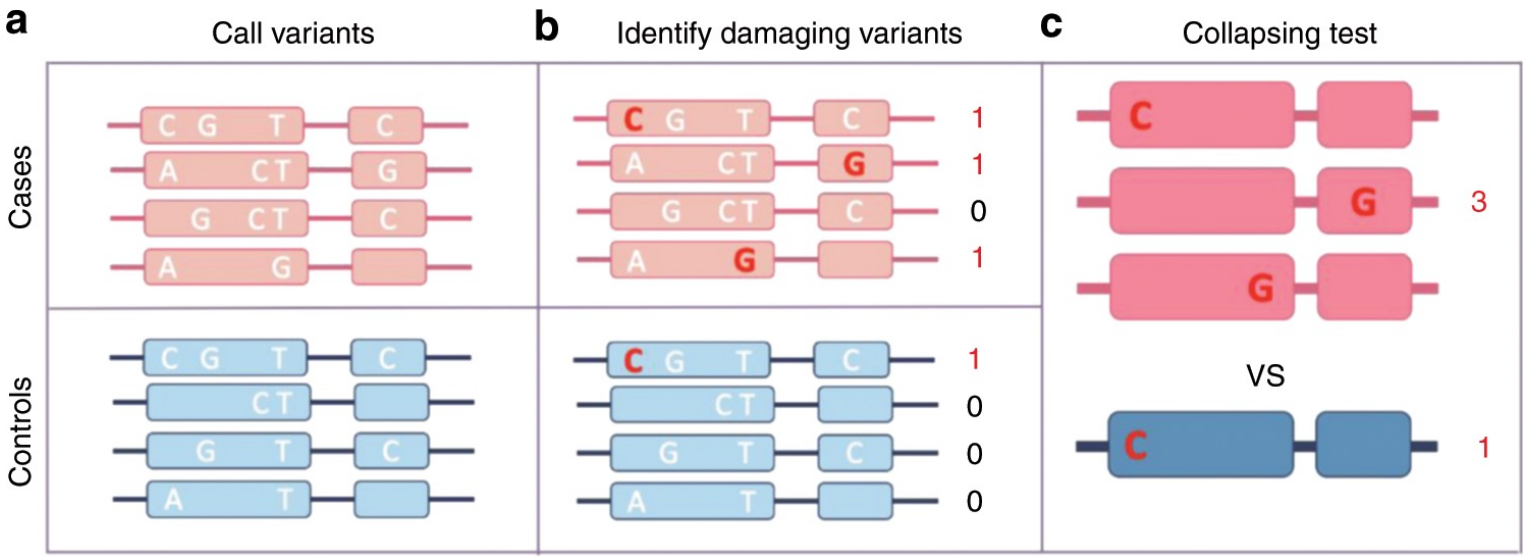
- Once these statistics are loaded, we can perform the analyses on-the-fly within the browser itself, rather than asking the server to compute our results
  - Allows web tools to use new methods on sensitive data that cannot be uploaded

- Possible analyses: rare variant aggregation tests, conditional analysis, colocalization…

Chen H, Huffman JE, Brody JA, Wang C, Lee S, Li Z, et al. Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. Am J Hum Genet. 2019;104: 260–274.

Rare variant aggregation tests

- Combine evidence from multiple rare variants in a gene (or region) to detect association with disease



Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun. 2020;11: 542.

LocusZoom: visually explore aggregation tests

In the future, LZ could display aggregation test results

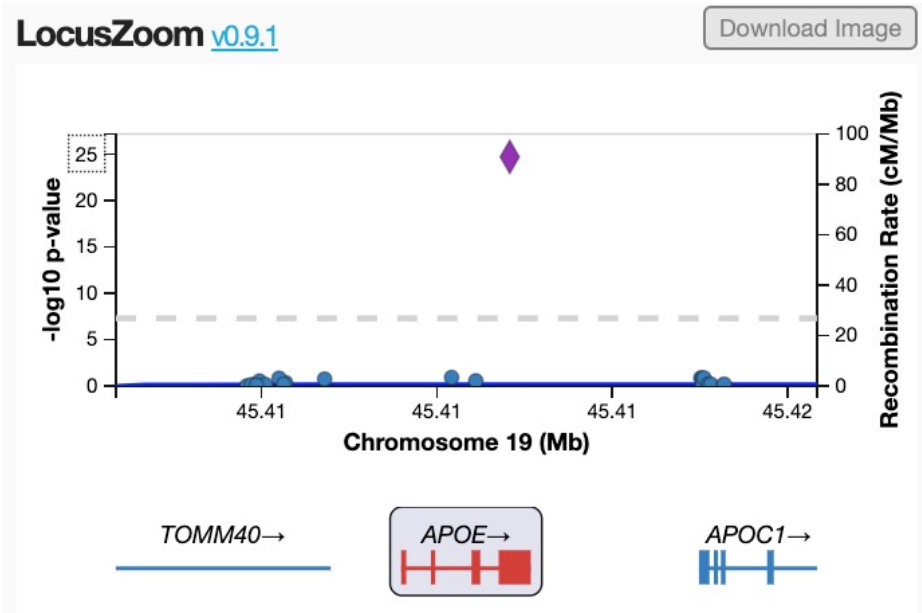


Table of aggregation test results

Currently reviewing results for group: (all genes in view)
Click on a row in the table below to see the variants associated with a specific mask + gene.

DOWNLOAD

Gene	Mask	# Variants	Test type	p-value
	filter column...		filter column...	
APOE	Exonic SNVs from GENCODE genes with allele frequency < 5%	3	skat	4.09 × 10^-24
APOE	Exonic SNVs from GENCODE genes with allele frequency < 5%	3	skat-o	8.16 × 10^-23
APOE	Exonic SNVs from GENCODE genes with allele frequency < 5%	3	burden	1.44 × 10^-19
APOE	Exonic SNVs from GENCODE genes with allele frequency < 5%	3	vt	4.31 × 10^-19
TOMM40	Exonic SNVs from GENCODE genes with allele frequency < 5%	12	skat	1.76 × 10^-5
TOMM40	Exonic SNVs from GENCODE genes with allele frequency < 5%	12	skat-o	2.47 × 10^-5
TOMM40	Exonic SNVs from GENCODE genes with allele frequency < 5%	12	vt	4.49 × 10^-4

Variants in mask Exonic SNVs from GENCODE genes with allele frequency < 5% / APOE

DOWNLOAD

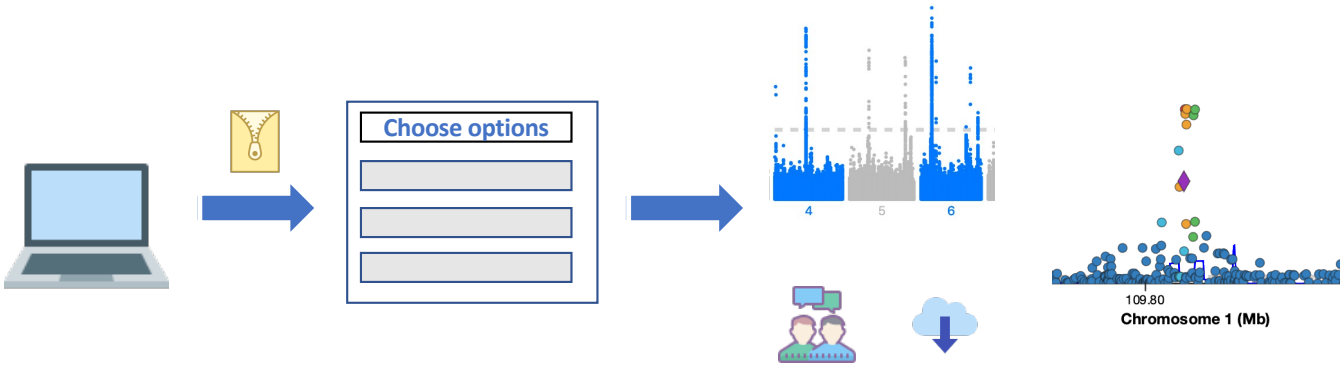
Variant	p-value	Alt allele frequency
19:45410420_G/A	0.122	0.002
19:45411110_T/C	0.280	0.008
19:45412079_C/T	1.82 × 10^-25	0.044

Balancing the old and the new

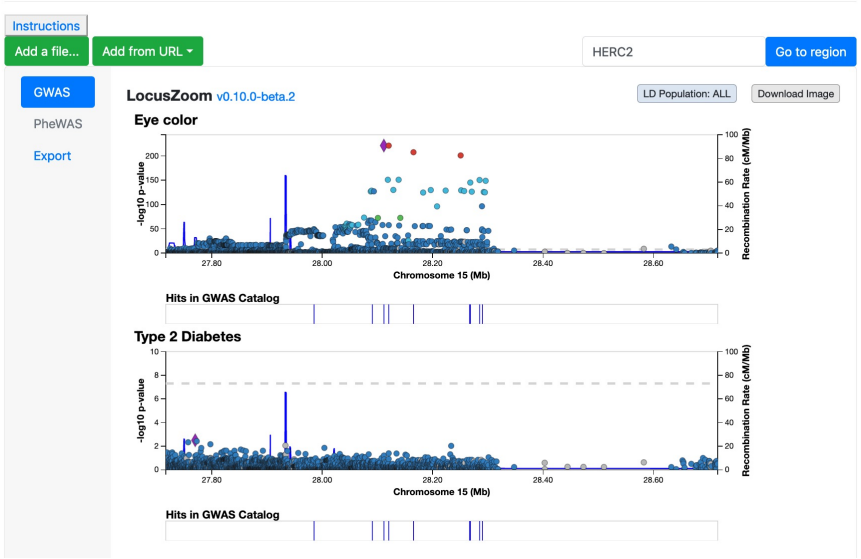
Providing user-focused tools to drive future adoption of LocusZoom.js

Making it easier to get started

- We are building new ways to use LocusZoom.js with user-provided data
- Serves as a testbed for new LocusZoom features
- Do what web browsers are good at, rather than copy the command line tool



LocalZoom: Quick Region Plots, without uploading

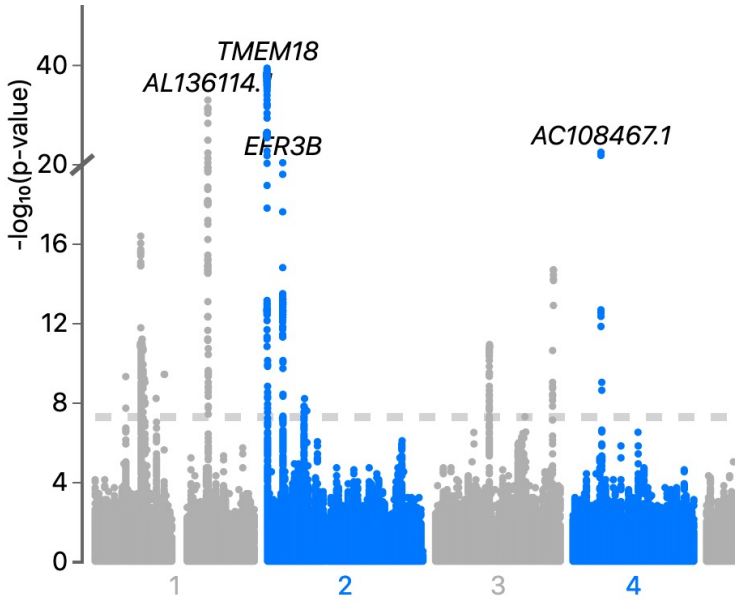


- Make a standard LocusZoom plot from tabixed GWAS summary statistics, in a web browser
  - No upload required
- Full support for build GRCh37 and GRCh38
- Automatic annotations
- Explore hits for any variant in a PheWAS of the UK BioBank

https://statgen.github.io/localzoom/

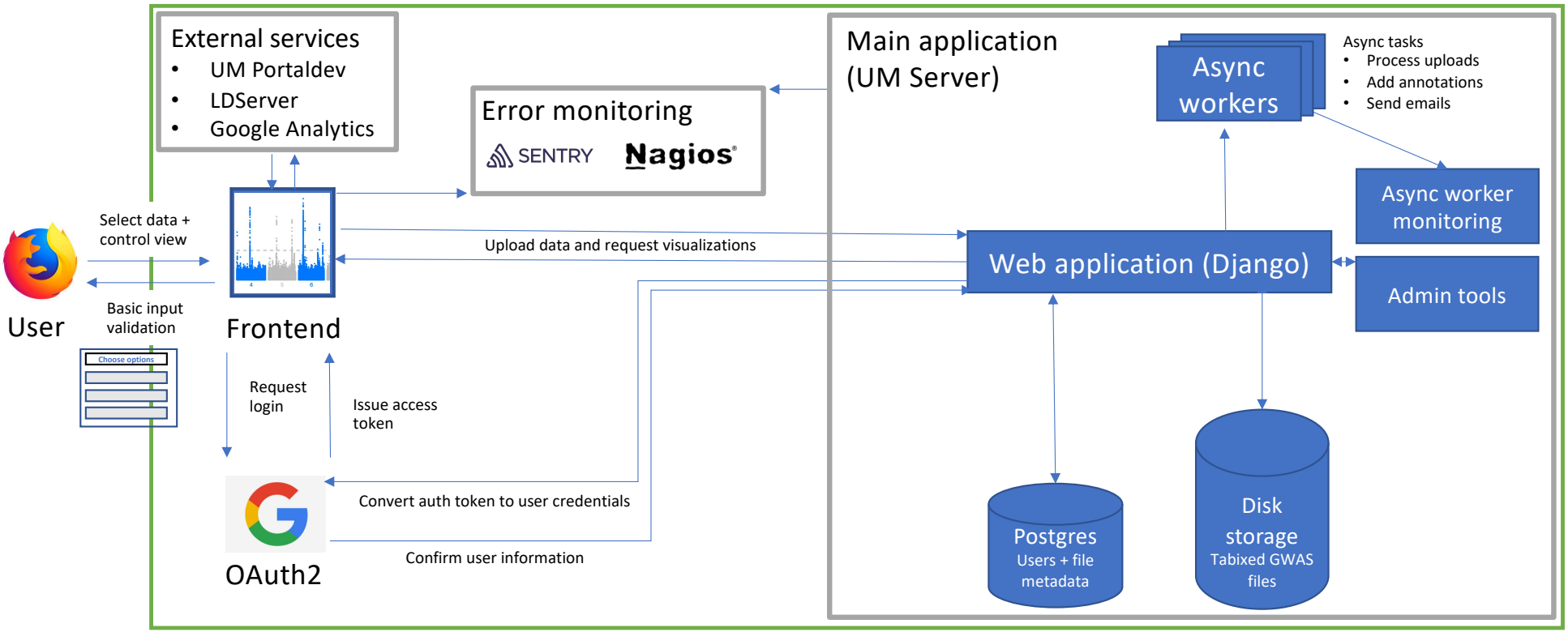
Upload, analyze, and share: my.locuszoom.org

- Enhanced view with additional information: top loci, Manhattan plot
- Supports a variety of GWAS file formats (no need to run tabix first)
- New website allows you to make your data public, or preview an analysis privately and share with trusted collaborators
- Explore publicly shared results from other studies



https://my.locuszoom.org/

my.locuszoom.org: Architecture



https://github.com/statgen/locuszoom-hosted/blob/develop/production.yml

“Upload your data”: Challenges and solutions

Challenge

- Must display data from many sources in a consistent way
- Users need to trust us with a copy of their data
- Features may require input from > 1 developer
- Shared server: a slow upload could reduce service for everyone
- We pay for all storage

Solution

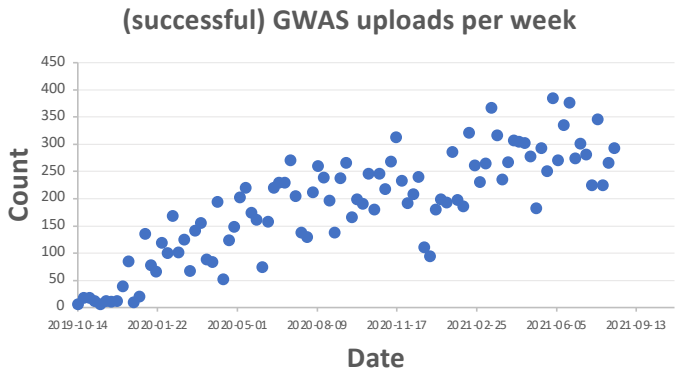
- Restrict allowed fields and provide aggressive validation
- Offer private sharing + option to delete; use well tested libraries
- Automate deployment (Docker) and DB updates (Django)
- Limit max file size; explore batch computing services
- Gonçalo buys more hard drives*

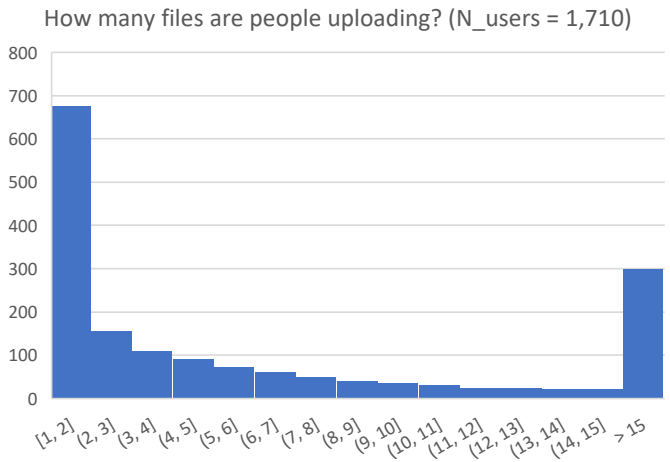
* When in doubt: “The best DevOps tool is MONEY”

Evaluation and Engineering

Offering tools at scale

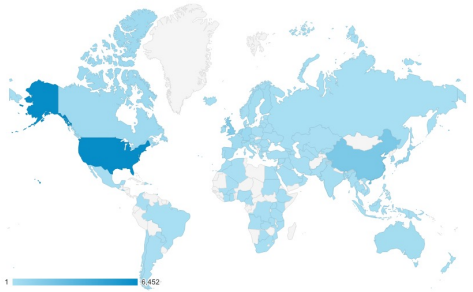
What is the measure of success?

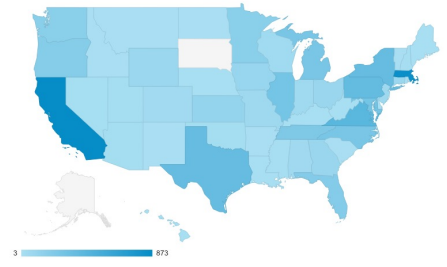




For https://my.locuszoom.org :

- Total user accounts: 3167
  - 1,278 returned > 7 days after signup
  - 48 US states and 108 countries/ territories
- 21,174 studies ingested and 3,279 ”private sharing links”

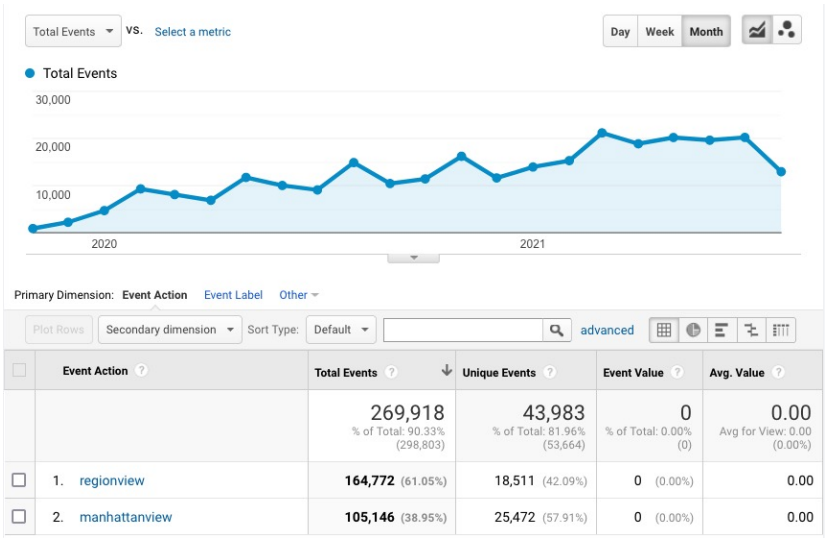




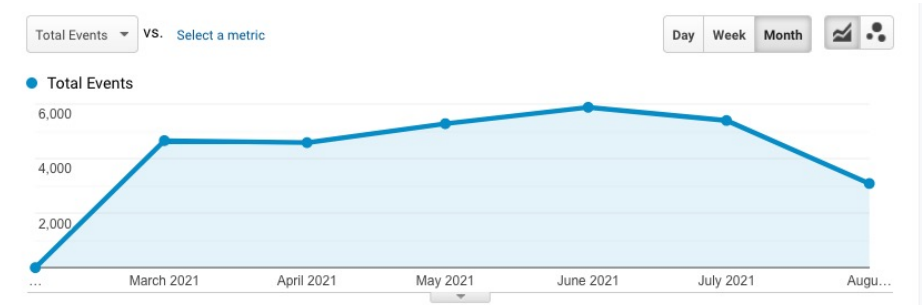
What Features Do People Like?

- Feature-level metrics let us see which things get used
- As we add new visualization types this will help make decisions about which buttons to keep- or remove.

Beyond Pageviews: analytics by feature/ event



	Event Action	Total Events	Unique Events	Event Value	Avg. Value
		269,918 % of Total: 90.33% (298,803)	43,983 % of Total: 81.96% (53,664)	0 % of Total: 0.00% (0)	0.00 Avg for View: 0.00 (0.00%)
1.	regionview	164,772 (61.05%)	18,511 (42.09%)	0 (0.00%)	0.00
2.	manhattanview	105,146 (38.95%)	25,472 (57.91%)	0 (0.00%)	0.00



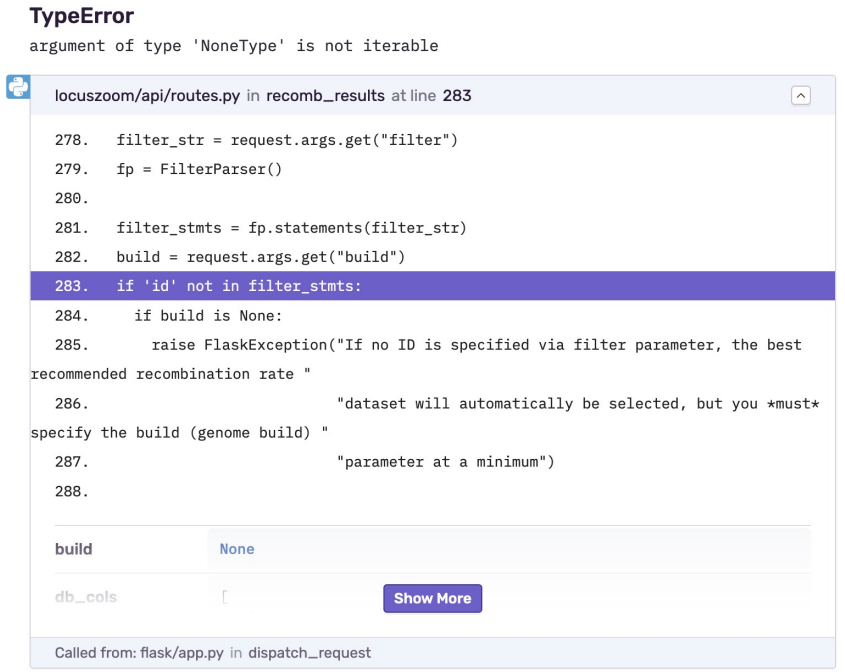
	Event Action	Total Events	Unique Events	Event Value	Avg. Value
		28,885 % of Total: 18.37% (157,256)	9,681 % of Total: 32.31% (29,962)	0 % of Total: 0.00% (0)	0.00 Avg for View: 0.00 (0.00%)
1.	save_image	7,823 (27.08%)	1,687 (17.43%)	0 (0.00%)	0.00
2.	ld_pop	7,031 (24.34%)	3,200 (33.05%)	0 (0.00%)	0.00
3.	association_display_options	6,464 (22.38%)	2,929 (30.26%)	0 (0.00%)	0.00
4.	ld_refvar	5,186 (17.95%)	852 (8.80%)	0 (0.00%)	0.00
5.	filter_genes	2,381 (8.24%)	1,013 (10.46%)	0 (0.00%)	0.00

Example: ~10% of region plots are saved for followup

Avg 10 region plots generated for every file uploaded

Web Dev: The Art of debugging someone else's computer

- Unlike command-line tools, you don't control where JavaScript runs
  - Even if a bug gets reported, it may not contain enough information to act on
- Use tools that automatically capture and report errors
  - Email when servers go down
  - Report browser stack trace with exact line of code where error occurred
  - Profile slow code on the server



SENTRY Nagios® DATADOG

Next steps: Science and engineering together

- Support more types of visualizations and connections between data
  - Gene Perturbation experiments
  - User-driven analysis (conditional analysis)
- Make deployments faster, more robust, and easier to recover
  - Our API server is now an incidental point of failure for 100+ sites that use LocusZoom.js
- Make it easier to "plot your own" data for unpublished studies, with a minimum of data munging
  - BED files
  - Custom user-provided LD
  - ...other requests?

Free, modular, reusable tools

Code, instructions, and examples are available for download

- LocusZoom.js: https://github.com/statgen/locuszoom
- Credible sets: https://github.com/statgen/gwas-credible-sets
- Aggregation tests: https://github.com/statgen/raremetal.js
- Michigan LDServer: https://github.com/statgen/LDServer
- PheWeb: https://github.com/statgen/pheweb
- my.locuszoom.org: https://github.com/statgen/locuszoom-hosted
- LocalZoom: https://github.com/statgen/localzoom/

Advice: Technology

- Plan for the technical limits of your system
  - Example: Python & JS use 64-bit IEEE754 floats, and very small pvalues cannot be represented (they underflow and are rendered as "0")
  - Store and transmit $-\log_{10}(p)$ values to avoid truncating the most interesting hits
- Benchmark performance for worst-case; expect datasets to grow
- Have a plan for maintenance
  - Use technologies that your team can agree on
  - Automate wherever possible: insist on reproducibility
  - Write unit tests to verify that your calculation methods work after a change
- Be pragmatic: add features in small chunks

Advice: Know your audience

- Offer a compelling service
  - Get interesting datasets
  - Know what your tool is trying to do: have a user in mind, and ask for feedback
  - Provide options, but measure which ones get used
- Metrics help guide design decisions- think beyond page views
- Enable independent use: FAQ, helpful error messages, and public issue tracker. Let people answer the common questions themselves

...And of course, contribute back to open source! Bug reports, documentation fixes, and feature requests are **valuable**!