Data Integration Case Study:

TERSECTBROWSER

Dr Tomasz Kurowski
t.j.kurowski@cranfield.ac.uk

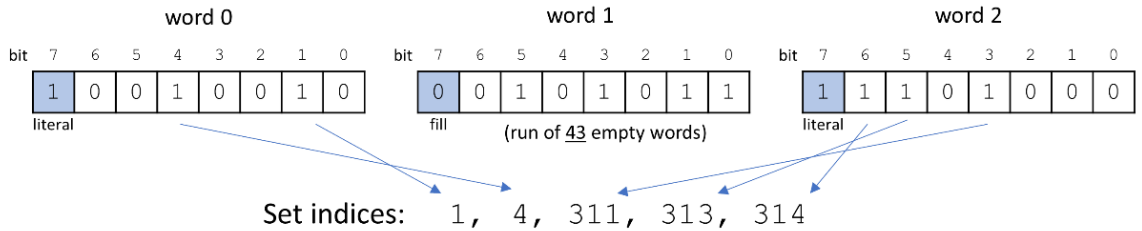
13 February 2025

www.cranfield.ac.uk

Tersect

Simple utility supporting a set theory syntax and bit array indices with Word-Aligned Hybrid lossless compression. Takes advantages of SSE processor extensions for speed.

A) Sample WAH-compressed bit array

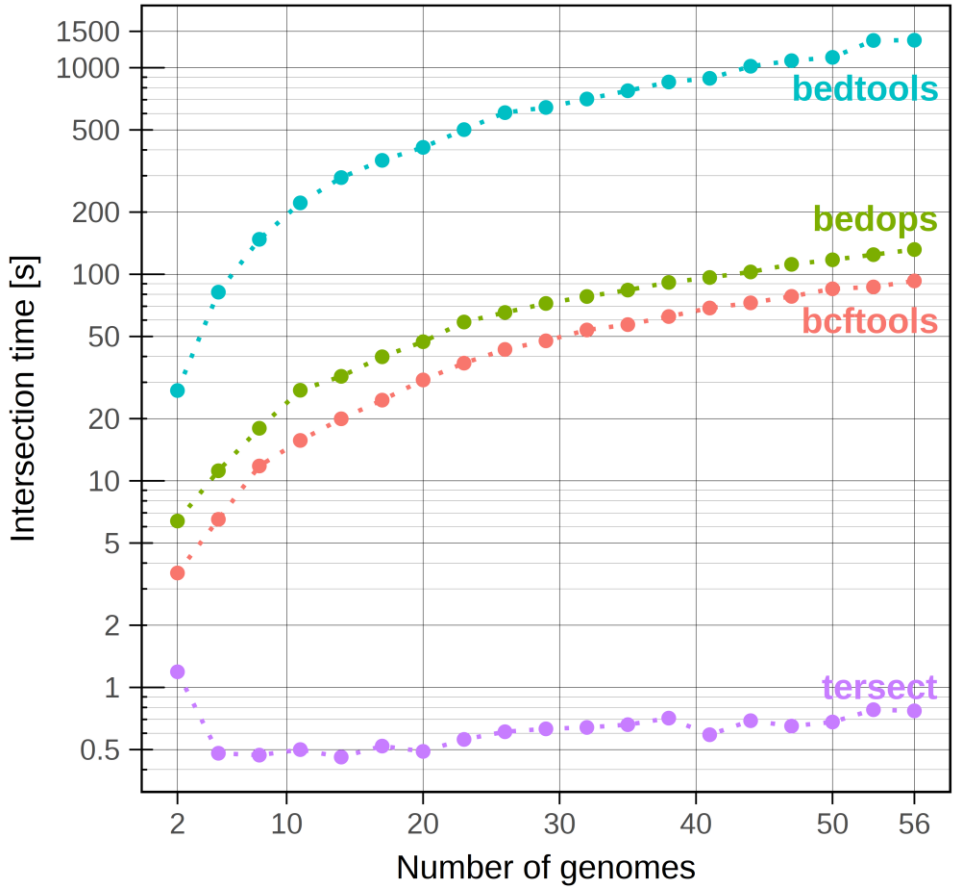


B) Chromosome variant list

index	position	data
0	22	G/T code
1	104	A/C code
2	159	C/G code
3	159	C/T code
4	1332	"GT/G" offset
5	4492	T/A code
...	...	...
309	95569	A/C code
310	96431	C/A code
311	96833	T/A code
312	97340	G/A code
313	104814	G/C code
314	104814	G/T code
315	105491	C/A code

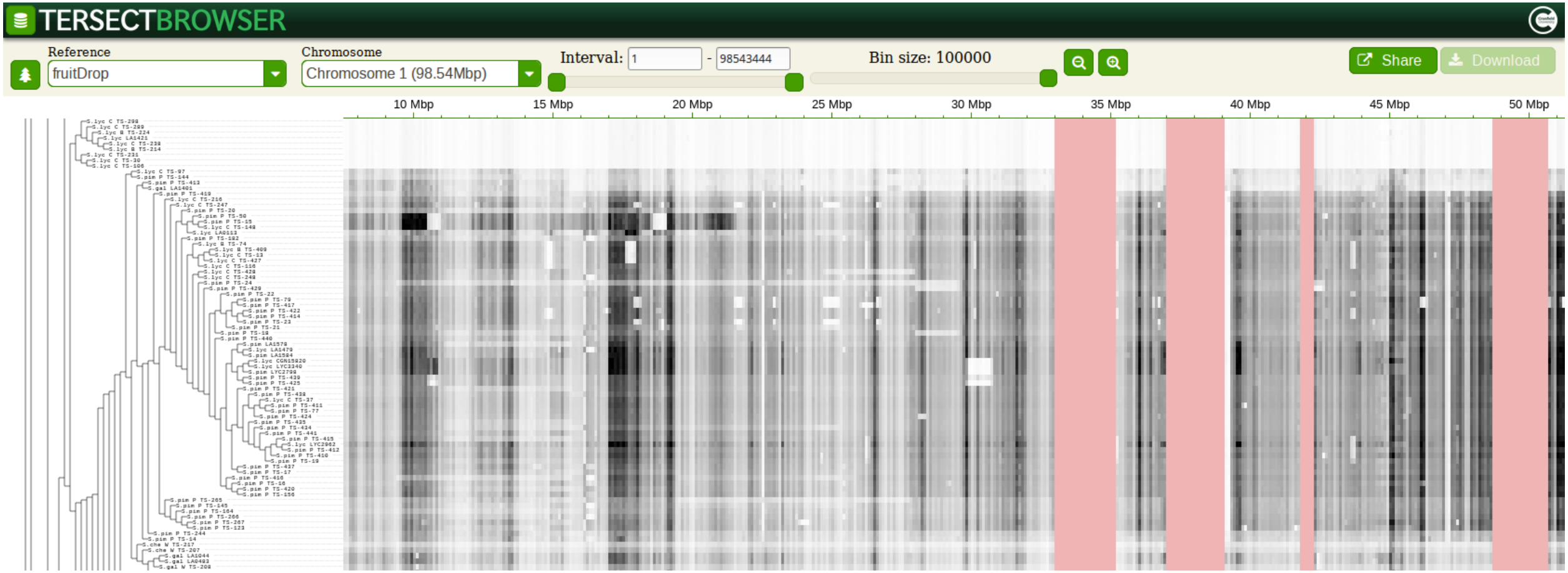
C) Sample variant output

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	104	.	G	T	.	.	.
chr1	1332	.	GT	G	.	.	.
chr1	96833	.	T	A	.	.	.
chr1	104814	.	G	C	.	.	.
chr1	104814	.	G	T	.	.	.



Kurowski TJ, Mohareb T. (2020) Tersect: a set theoretical utility for exploring sequence variant data. Bioinformatics, Volume 36, Issue 3, February 2020

Tersect Browser



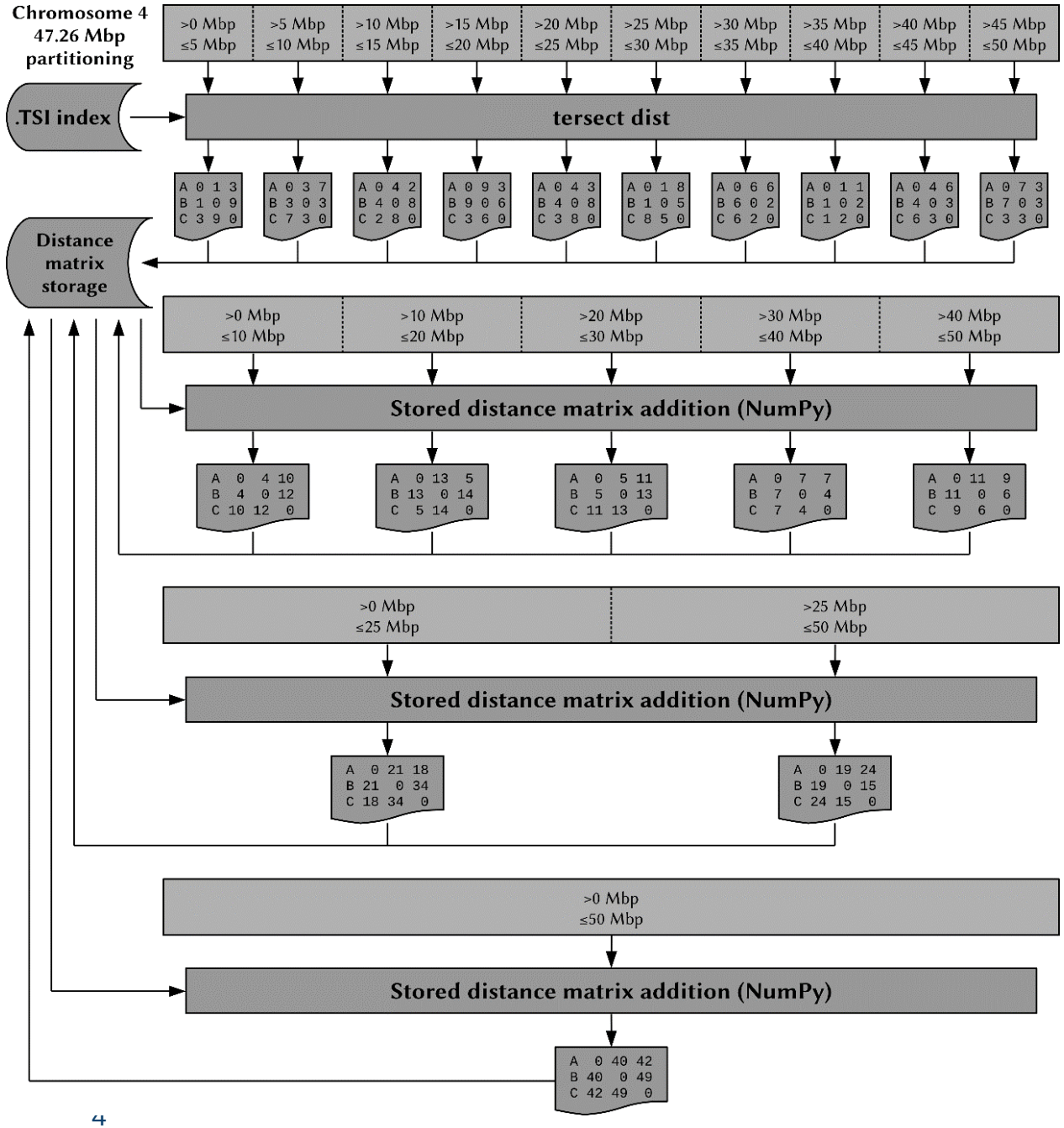
Interface using Angular (and exploiting CSS acceleration...)

Precalculation

Some results can be pre-calculated and later adjusted to specific queries.

(Indexing is a type of pre-calculation too...)

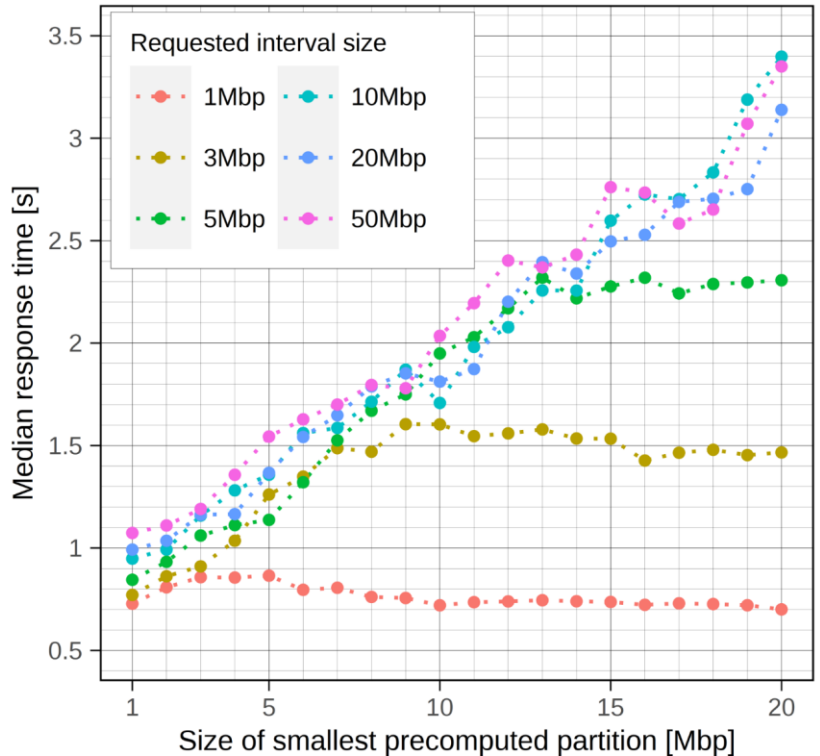
Better to store this in a format that is **easy to manipulate**, instead of anything like the final result.

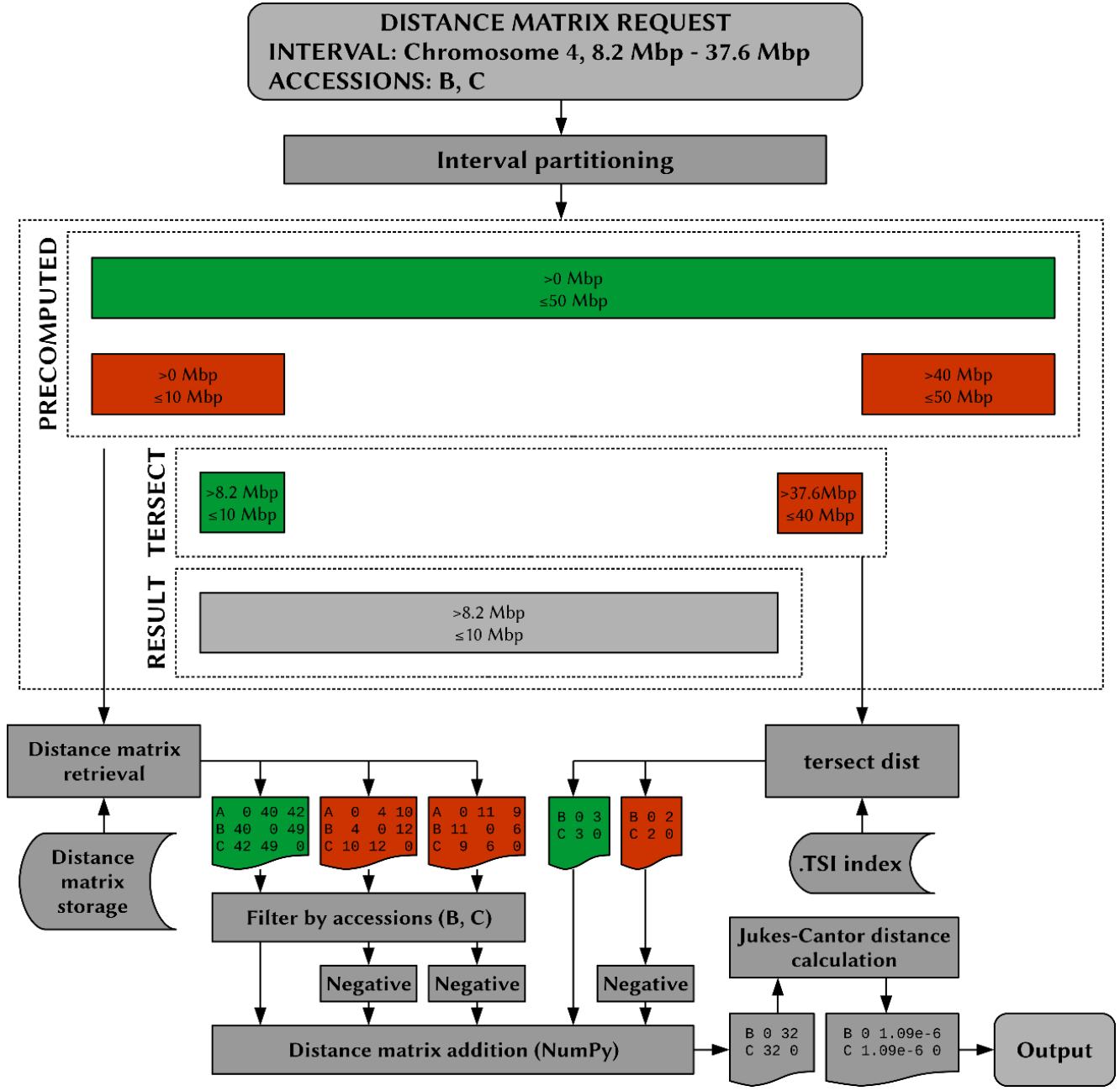


Retrieval

With precalculated results for each 5Mbp interval, I never have to generate new results for more than 5Mbp.

By setting the interval size, I can tune the performance speed (costs storage).

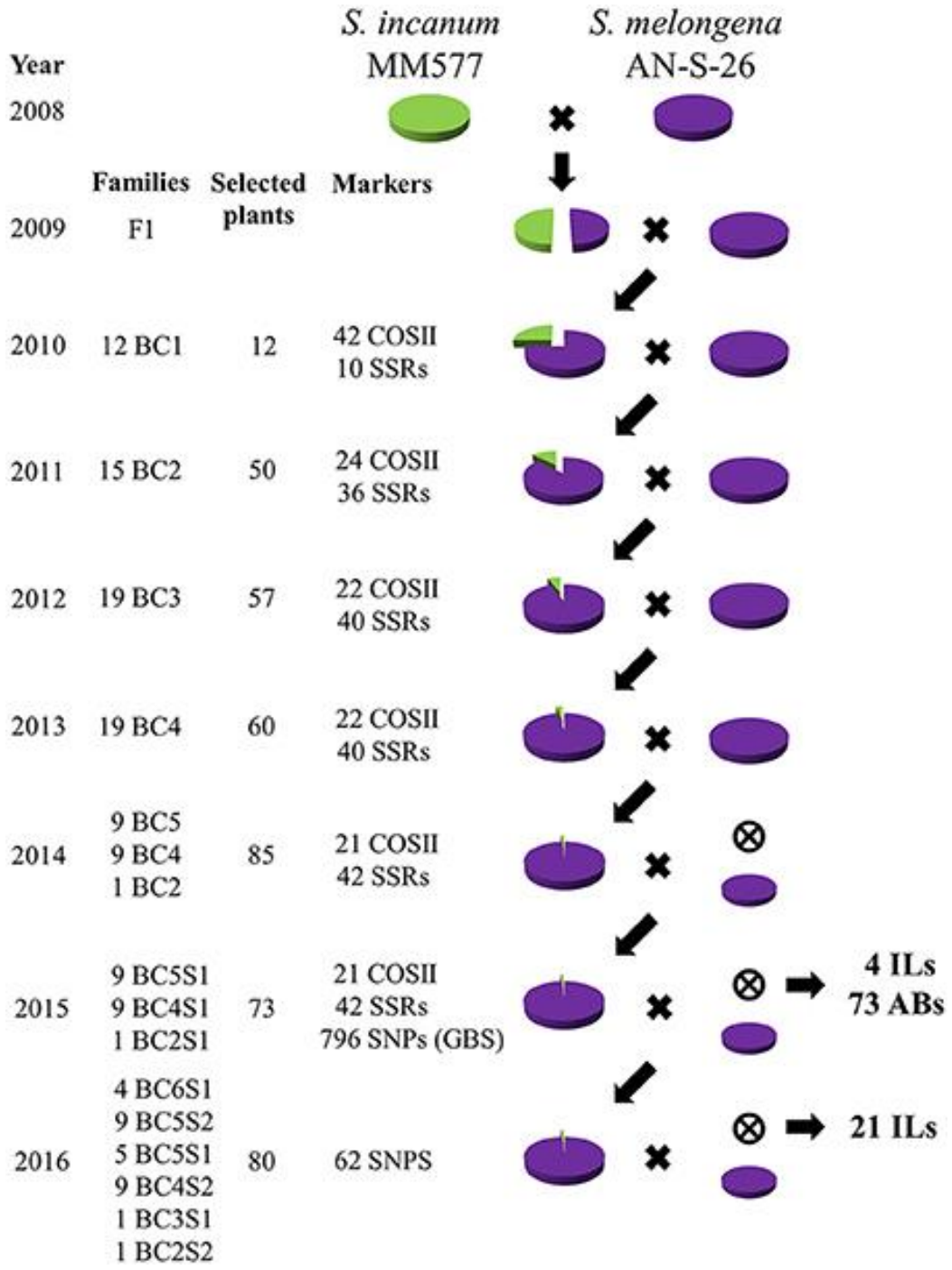




Introgressive hybridisation

You can introduce genes from (even across species) into elite cultivars through repeated back-crossing and selection.

In theory you could narrow down your introgression to a single gene.



Non-model organism resources

Tomato genome sequence builds

Release	Date	Description	Annotation	Download
SL1.00	Dec 2009	Initial build, based on the Newbler assembler and containing only 454 sequencing data.	ITAG1	scaffolds proteins cds
SL1.03	Jan 2010	Like 1.00, but with additional 454 runs and improved contamination screen.	Not annotated	scaffolds
cabog1.00	Mar 2010	All 454 data, bac end and fosmid end data, assembled using the CABOG assembler.	Not annotated	scaffolds
SL1.50	Apr 2010	Includes all 454 data, bac ends, fosmid ends, polishing with Solexa and SOLiD data.	Not annotated	scaffolds
SL2.00	Jun 2010	Release withdrawn.	Not annotated	-
SL2.10	Jun 2010	Additional scaffold merging using clone end sequences. Scaffolds placed and oriented using multiple physical maps, first release to include chromosome pseudomolecule sequences.	Not annotated	scaffolds, chromosomes
SL2.30	Aug 2010	Integration and polishing of tomato BAC sequences	moved to SL2.31	scaffolds, chromosomes
SL2.31	Nov 2010	Mask a small number of contaminated regions. Base-compatible with SL2.30.	ITAG2	scaffolds, chromosomes
SL2.40	Jan 2011	Small amount of additional contamination removal. Regularize gap sizes to comply with GenBank policies.	ITAG2.3	scaffolds, chromosomes
SL2.50	Feb 2014	Rearrangement of scaffolds and a number of gaps re-sized according to FISH data.	ITAG2.4	scaffolds, chromosomes
SL3.00	Feb 2017	BAC integration, Chr00 integration and BioNano data	ITAG3.20	chromosomes
SL4.00	Sept 2019	*De novo* assembled PacBio genome scaffolded with Hi-C. Validated using Bionano and 10X linked-reads	ITAG4.0	chromosomes

You may need to build your own reference genome assembly, and those are often a „work in progress”.

GENOME DATABASE FOR ROSACEAE

Resources for Rosaceae Research Discovery and Crop Improvement

Rosaceae: https://www.rosaceae.org/

Sol Genomics Network

Solanaceae: https://solgenomics.net/

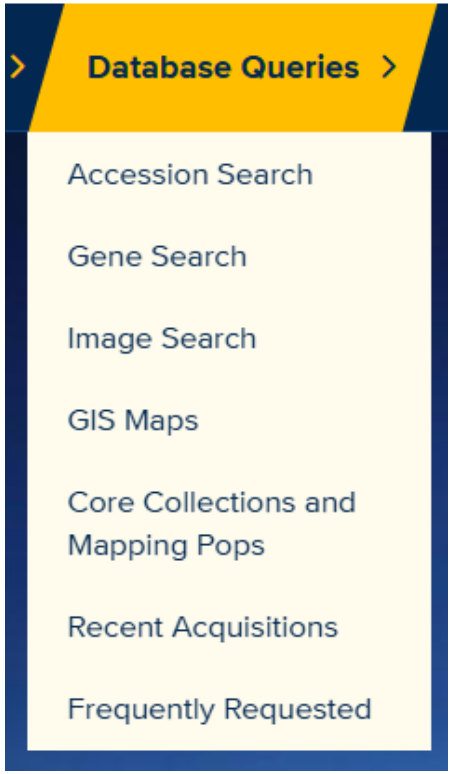
TGRC

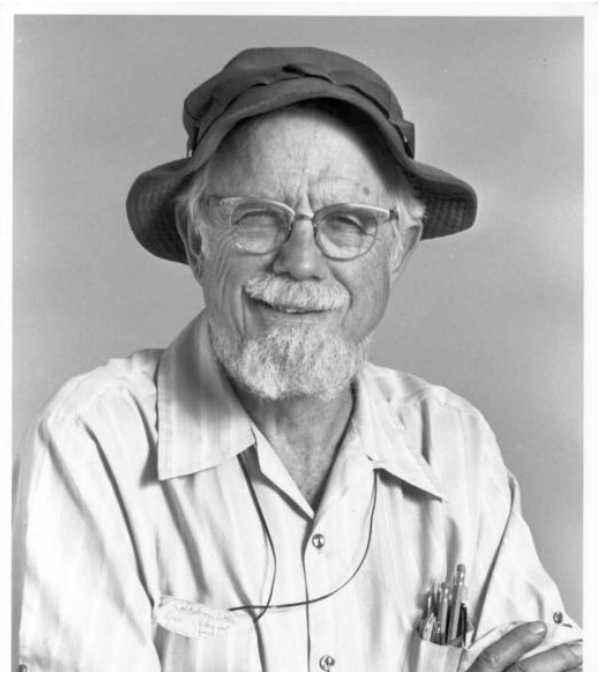
~OMIM for tomatoes

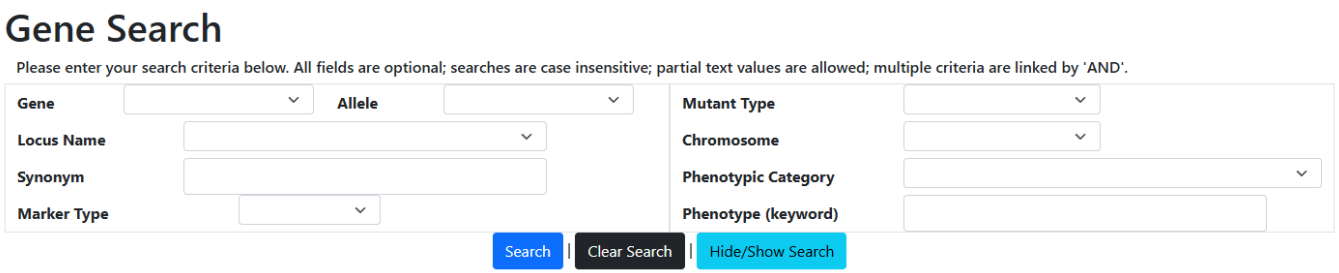
Maintained by UC Davis.

No (public?) programmatic access!

I had to create my own copy...







C.M. Rick

TGRC

Tomato Genetics Resource Center

Web scraping

I wrote a bot which scraped the TGRC database using the Requests and Beautiful Soup modules.

Then I made a REST API serving the data to Tersect Browser.

[Download | Documentation | Hall of Fame | For enterprise | Source | Changelog | Discussion group | Zine]

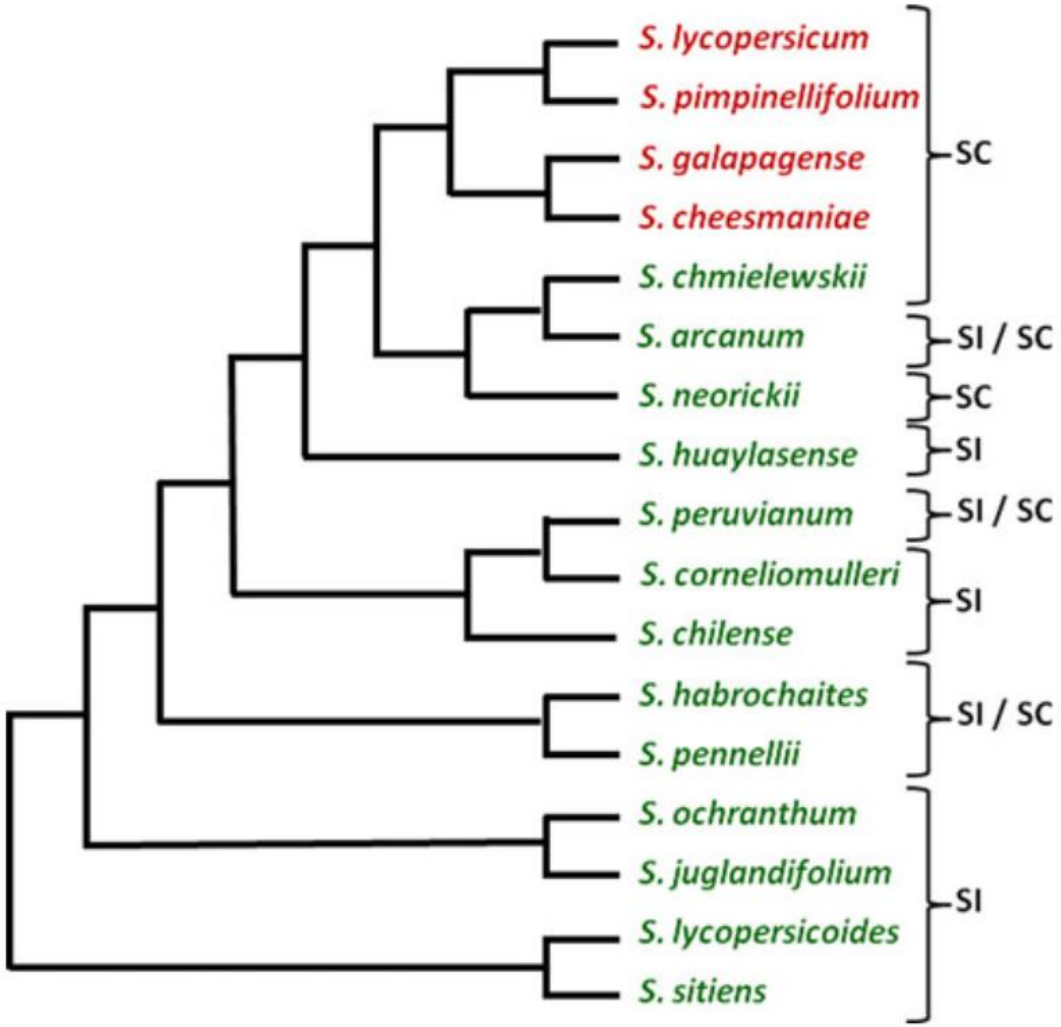
Beautiful Soup

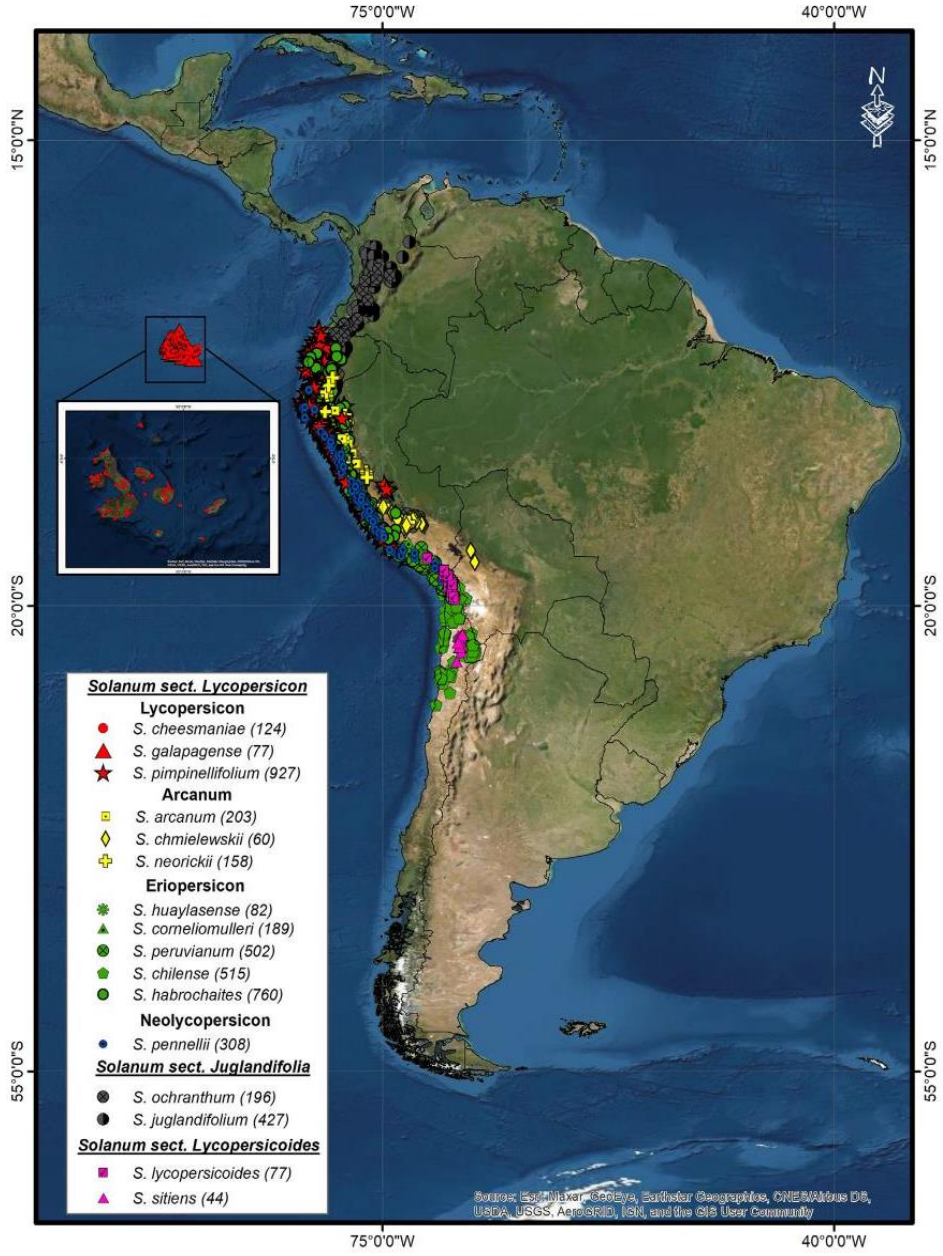
You didn't write that awful page. You're just trying to get some data out of it. Beautiful Soup is here to help. Since 2004, it's been saving programmers hours or days of work on quick-turnaround screen scraping projects.

Beautiful Soup is a Python library designed for

```
router.route('/accessions/:gene?/:filter?')
    .get((req, res) => {
    const gene = req.params.gene;
    const filter = req.params.filter || false;
    const query = gene ? { 'alleles.gene': gene } : {};
    const projection = { _id: 0, accession: 1, alleles: 1};
    AccessionTGRC.find(AccessionTGRC.translateAliases(query),
                       AccessionTGRC.translateAliases(projection))
                 .exec((err, result: AccessionTGRC[]) => {
        if (err) {
            return res.status(500).send('Accessions could not be retrieved');
        } else {
            const output = result.map(acc => {
                const accObj = acc.toObject();
                return {
                    accession: accObj.accession,
                    alleles: accObj.alleles
                };
            });
            if (filter) {
                // Exclude other genes from result
                output.forEach(acc => {
                    acc.alleles = acc.alleles.filter(a => a.gene === gene);
                });
            }
            return res.json(output);
        }
    });
});
```

Tomato and its wild relatives





Inflorescence control

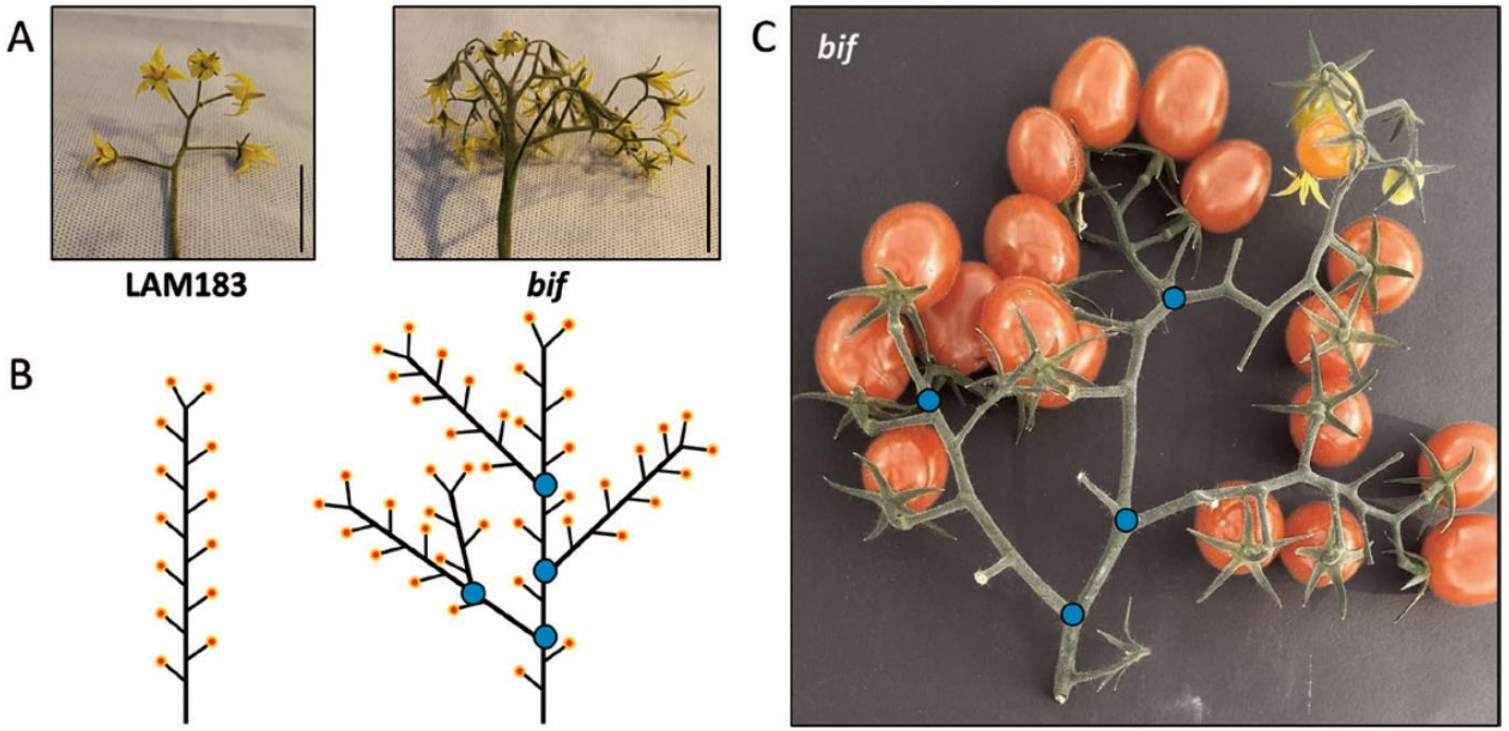
SlMAPK1 gene mutant null allele associated with *bif* phenotype identified on chromosome 12.

Its origin was determined as a ~2Mbp introgression from *Solanum galapagense*, a wild tomato species native to the Galápagos Islands.

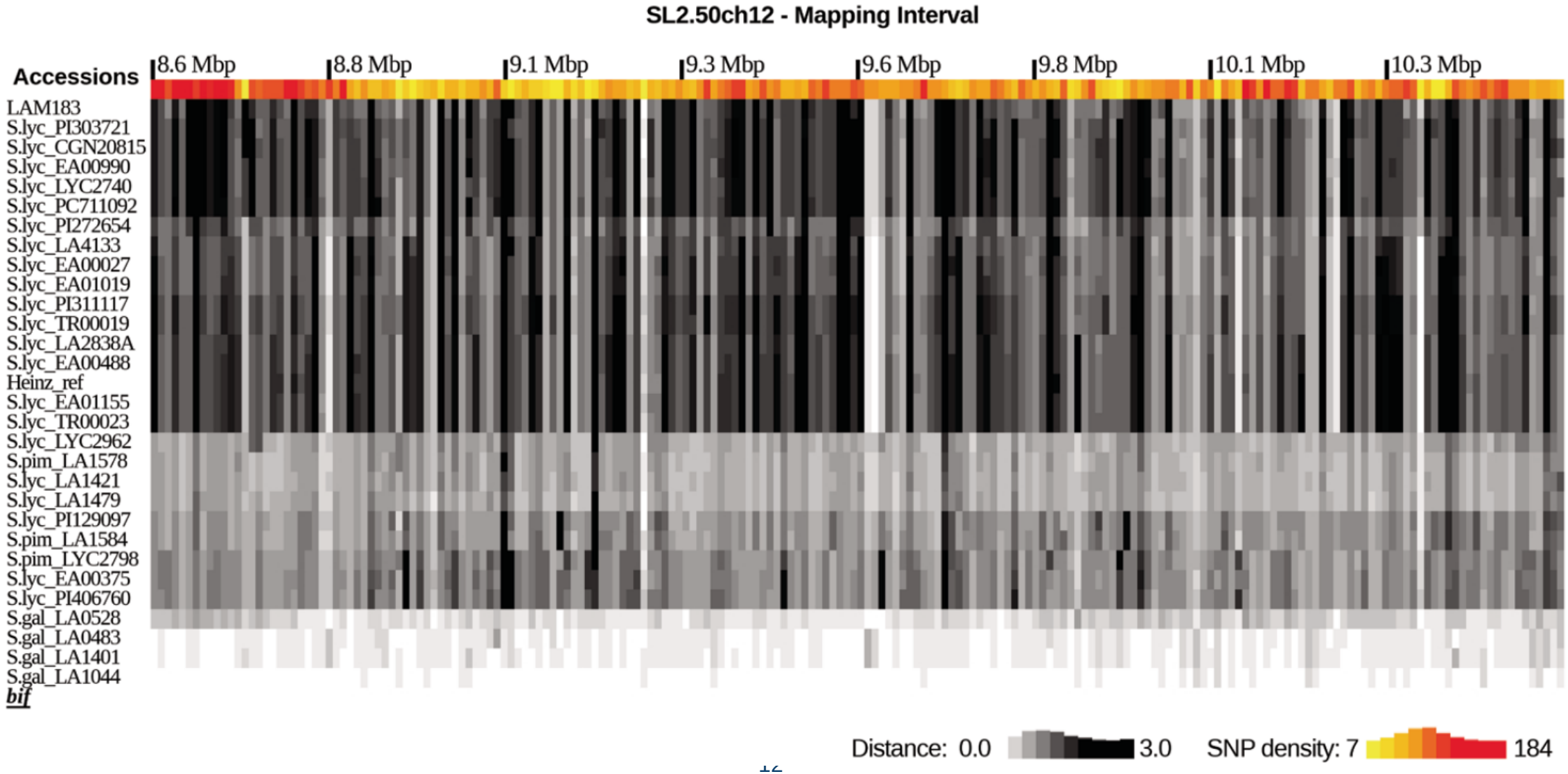
BIFURCATE FLOWER TRUSS: a novel locus controlling inflorescence branching in tomato contains a defective MAP kinase gene

Demetryus Silva Ferreira, Zoltan Kevei, Tomasz Kurowski, Maria Esther de Noronha Fonseca, Fady Mohareb, Leonardo S Boiteux, Andrew J Thompson

Journal of Experimental Botany, Volume 69, Issue 10, 27 April 2018, Pages 2581–2593, https://doi.org/10.1093/jxb/ery076



Galapagense introgression



Working on BBSRC bid to expand the tool further!

There may be projects based on this!